Texas Success Initiative Assessment 2.0 Technical Manual

Characteristics of TSIA2

College Board

June 2021

Foreword

The Texas Success Initiative Assessment 2.0 (TSIA2) is a series of placement and diagnostic tests for students enrolling in public colleges and universities in Texas. The tests help Texas schools determine whether students are ready for college-level courses in the areas of reading, writing, and mathematics. As the vendor who develops and administers the TSIA2, College Board understands that the challenge of getting students to and through college is a shared responsibility. We are honored to have the opportunity to work with educators, institutions, and policymakers to make sure that the best information is available about what Texas students know and can do so that accurate and fair course placement decisions can be made, and students can be offered the support services they need to succeed.

The TSIA2 includes reading, writing and mathematics assessments that determine if an incoming student is prepared to enroll and succeed in entry-level college courses or if additional skill development is needed. TSIA2 provides a detailed analysis of the test taker's strengths and weaknesses to help focus support on the identified areas of development with the goal of skill building and eventual mastery. TSIA2 includes questions for classification and diagnostic purposes that align with Texas College and Career Readiness Standards (CCRS); the critical Texas Essential Knowledge and Skills (TEKS) and CCRS Performance Expectations that support the English III (Reading and Writing) and Algebra II State of Texas Assessments of Academic Readiness (STAAR) End-of-Course Assessments; the AEL Standards 2.0; and the skills identified under the National Reporting System's (NRS) six-level Educational Functioning Level Descriptors (EFLD).

This technical manual documents the processes of the design, administration, and scoring of the TSIA2 assessments. More importantly, this manual represents the base of evidence supporting the development and psychometric quality of the assessments, including reliability, validity, fairness of use, and standard setting. Research-based evidence is the hallmark of College Board's work, and we are pleased to present this comprehensive documentation in accordance with professional testing standards. As an online, internet-based computer-adaptive test, TSIA2 represents an innovative and convenient tool to aid Texas educators in advancing college readiness and success for Texas students. College Board is proud to be the Texas Higher Education Coordinating Board's partner in providing this service.

Denny Way, Ph.D.

Senior Advisor to the SVP, Learning, Evaluation and Research

College Board

Contributors

The following individuals were involved in the creation of the *TSIA2 Technical Manual.* We thank them for sharing so generously of their time, effort, and expertise.

Chapter Leads

Luz Bay, Psychometrics

Constance Tsai, Assessment Design and Development

James Chandler, Group Lead

Writers/Reviewers

Siang Chee Chuah, Alex DeFazio, Thomas Gleiber, Lauren Molin, Jim Patterson, Lei Wan, Caiyan Zhang

Technical Manual Working Group

Mark Syp, Chief Editor

Luz Bay, Psychometric Lead

James Chandler, Group Lead

Constance Tsai, Assessment Design and Development Lead

Leadership Reviewers

Emily Paulsen, Executive Director of ACCUPLACER

Sherral Miller, Vice President of Assessment Design and Development

Thomas Proctor, Executive Director of Psychometric Analytics and Digital Efforts

David Williamson, Vice President of Psychometrics

Denny Way, Senior Advisor to the SVP, Learning, Evaluation and Research

Special Thanks

Nelson Chow

Dara Meiner

Table of Contents

List of Tables

List of Figures

Preface

Purpose of Manual

The purpose of this manual is to provide information about the technical qualities of the tests that make up the Texas Success Initiative Assessment 2.0 (TSIA2). These tests consist of both a College Readiness Classification (CRC) Test and a Diagnostic Test in two subject areas, English language arts and reading (ELAR) and mathematics, as well as the Essay Test. This manual discusses the purpose of the tests and the rationale and principles behind their design and development. It describes the content of the tests; the procedures and processes that were and continue to be undertaken in the design, development, administration, and scoring of the tests; the appropriate interpretation of results; the consistency of the scores from a measurement perspective; and evidence that bears on the validity of interpretations made on the basis of the scores.

College Board believes that it is essential to provide documentation of this nature, in keeping with our organization's commitment to transparency and our desire to adhere to industry best practices and the AERA/APA/NCME standards governing supporting documentation for tests (found in Chapter 7 of the 2014 AERA/APA/NCME *Standards for Educational and Psychological Testing*). Maintaining assessments with strong evidence of validity supporting them is an ongoing process and will continue to evolve as the tests are administered. To that end, the information found in this technical manual is accurate as of the January 2021 launch of TSIA2.

A Quick Stylistic Note

The contents of this manual discuss the current iteration of the Texas Success Initiative Assessment, first administered in January 2021. Going forward, we will sometimes abbreviate this specific version of the test as "TSIA2." Sometimes, for the sake of comparison, it will be necessary to refer to the iteration of the test that was administered prior to 2021. In order to alleviate confusion, we will refer to the pre-2021 version as "TSIA1," even though the assessment did not actually bear this title while it was being administered. The abbreviation "TSIA" (with no number) will be used to refer to the overall Texas Success Initiative Assessment, not to a specific iteration of its tests.

Manual Contents

For ease of reading and understanding, this manual is structured to reflect the lifecycle of the TSIA2 tests. It provides insights about the tests, from their design through development, administration, scoring, and the interpretation of scores for intended uses.

As its name implies, Chapter 1: Overview provides an overview of the TSIA2 tests and their components and intended uses, including comparisons with components of the TSIA1 tests. Chapter 2: Fairness provides an examination of how College Board ensures that TSIA2 test content is inclusive, representative, and accessible, consistent with the designs and aims of the tests; the steps taken to

ensure tests are administered in a manner that is fair and equitable for all test takers; and the efforts made to support fair and valid interpretation of test scores.

Chapter 3: Test Development Procedures details the processes used to develop the TSIA2 tests. This includes information on the establishment of test specifications for each test in the suite; the test development process; the creation, review, and analysis of the questions that make up the tests; the ACCUPLACER computer-adaptive test (CAT) algorithm; and special accommodations for test takers who require alternative formats.

Marking a shift away from the test development portion of the manual, Chapter 4: Administration of TSIA2 reflects the next stage in the lifecycle of the tests, describing the procedures used and the security measures taken to administer them in a manner that supports their fair and valid use.

Following test administration, the focus of the manual shifts to the scores that are produced. Chapter 5: Interpretation and Application of Results looks at the scoring procedures and analyses used to ensure that placement decisions made using TSIA2 scores lead to correct conclusions. Chapter 6: Psychometrics takes this discussion one step further, as it describes procedures necessary to ensure that TSIA2 tests are the best possible assessment of that which they are intended to measure.

This manual concludes with Chapter 7: Validity, where the guiding principles of validity are presented. As this manual aims to make apparent, validity considerations permeate every aspect of TSIA2. We have chosen to discuss validity at this point in the manual, as validity evidence takes the results of all our previous analyses and addresses whether the tests can be used to determine college readiness, the overall goal of TSIA2.

Chapter 1 — Overview

Introduction

This chapter discusses the components that comprise the Texas Success Initiative Assessment 2.0 (TSIA2), as well as its intended uses. Section 1.1 provides an overview of the various TSIA2 tests with a look at their key features, including how they differ from those of the TSIA1 assessment. Section 1.2 briefly discusses the history of the assessment redesign, including the rationale behind the new components of the test and what it seeks to emphasize. Section 1.3 provides an overview of the content of the tests and their question and task formats, in keeping with standards and best practices.

1.1 Description and Features of the Assessment

Brief Description of Texas Success Initiative Assessment 2.0

The Texas Success Initiative Assessment 2.0 (TSIA2) is a revision of the TSIA1 designed, developed, and maintained by College Board on the ACCUPLACER® platform. In 2012, College Board entered into a contract with the Texas Higher Education Coordinating Board (THECB) to create and support TSIA1, with the goal of improving student success rates in Texas colleges. In 2019, College Board was awarded a contract to create an updated version of TSIA1 to serve that goal going forward.

TSIA2 contains several computer-adaptive multiple-choice tests as well as an essay test. Unless exempt, entering college students are required to take TSIA2 tests in English language arts and reading (ELAR) and/or mathematics that either certify them as college ready or provide a diagnosis to 1) facilitate entry into the appropriate developmental education course or 2) support co-enrollment in a developmental education course and an entry-level, credit-bearing course within the same semester.

Features of TSIA2

TSIA2 includes two multiple-choice College Readiness Classification (CRC) Tests—one for ELAR and one for mathematics—and corresponding Diagnostic Tests. All multiple-choice tests are computer-adaptive. A computer-adaptive test is a form of computer-administered test in which the subsequent question selected to be administered depends on the test taker's ability estimate based on the responses to the preceding questions.[1] The Essay Test, which remains unchanged from TSIA1, continues to be administered on computer. Accommodated forms continue to be available for each test.

College Readiness Classification (CRC) Tests. The CRC Tests correspond with the placement tests in TSIA1. All students, unless exempt, begin testing with the ELAR or Mathematics CRC Test. Students' CRC performance identifies them as either college ready or not college ready. Students who are identified as college ready can go on to enroll in any entry-level, credit-bearing college course in corresponding

[1] For more information on the computer-adaptive test engine that drives TSIA2 multiple-choice tests, see Chapter 5: Interpretation and Application of Results.

subject areas without restrictions or prerequisites, while those identified as not college ready are automatically administered the corresponding Diagnostic Test. The score reporting for the TSIA2 CRC Tests is on an 81-point scale that ranges from 910 to 990.

The TSIA2 Essay Test, which remains unchanged from TSIA1, is a direct assessment of writing designed primarily to ascertain (in conjunction with the ELAR CRC Test and possibly the ELAR Diagnostic Test) whether test takers are college ready or not college ready with respect to writing. Student essays are electronically scored on a scale from 1 to 8.

Diagnostic Tests. The TSIA2 suite contains two diagnostic tests, one for ELAR and one for mathematics. Each test incorporates TSIA1's standalone Adult Basic Education (ABE) testing components and covers the range of content and difficulty addressed by the previously separate TSIA1 diagnostic and ABE tests.

The TSIA2 Diagnostic Tests provide test takers with actionable information about their academic strengths and weaknesses across a range of content-based strands (reporting categories that yield content-based subscores) so that targeted instruction and intervention may be delivered. Each test yields a diagnostic profile that includes:

- a proficiency level (i.e., Basic, Proficient, or Advanced) along with proficiency statements on each of the test strands, and
- a classification into one of five diagnostic levels closely aligned to the National Reporting System Educational Functioning Levels (NRS EFL).

In addition, each Diagnostic Test yields a Learning Locator Code (LLC) that matches the test taker by achievement, based on performance on the diagnostic strands, to appropriate learning activities and materials available in a free learning tool called TSIA2 Learning Resources.[2]

Compared to TSIA1, TSIA2 offers a more integrated testing experience. In TSIA2, testing outcomes are reduced from TSIA1's two classification scores and three scoring categories (college ready, developmental education, adult basic education [ABE]) to a single college readiness classification score and two scoring categories (college ready, not college ready). Test takers who are not college ready are routed to diagnostic testing within the same testing experience. The Diagnostic Tests themselves combine the former Developmental Education (DE) Diagnostic and ABE tests. Finally, the separate multiple-choice Reading and Writing components from TSIA1 are combined into a single ELAR domain, further supporting the integrated testing experience of TSIA2. In addition to these adjustments, the TSIA2 has several additional new features, which are enumerated in the following paragraphs.

Connection to Texas curriculum and standards. One of the foundations of TSIA2 is a connection to classroom learning and experience. TSIA2 is designed to assess whether and to what extent students

[2] Delivered on Pearson's Perspective™ platform, TSIA2 Learning Resources provides supplemental learning materials and activities that students can access 1) pre-assessment, searching by topic, or 2) post-diagnostic testing, utilizing the individual LLC assigned, so that materials presented will address their knowledge and skill level.

have acquired the critical knowledge and skills that truly matter for success in college and career. Specifically, TSIA2 is deeply informed by and aligned to Texas's own academic and adult education literacy standards, as defined in the following:

- *Texas College and Career Readiness Standards (TXCCRS)*, which articulate the knowledge and skills that students must know and be able to apply to succeed in entry-level college/university courses and in the skilled workforce;
- *Texas Essential Knowledge and Skills (TEKS)* for English III and Algebra II**,** which indicate that a student is college ready and does not need to be enrolled in remedial or developmental education courses/interventions;
- *Texas Adult Education and Literacy Content Standards 2.0 (AEL 2.0)*, which outline the knowledge, skills, and abilities required for success at in-demand entry- and intermediate-level jobs in occupations within four industry clusters: advanced manufacturing; construction and extraction; healthcare sciences; and transportation, distribution, and logistics.

Connection to mathematics pathways. The TSIA2 Mathematics Tests are designed to help put students on a path to productive engagement in a society and economy that is increasingly reliant on data and quantitative reasoning. To achieve this goal, in both the TSIA2 Mathematics CRC and Diagnostic Tests, 1) quantitative reasoning constitutes its own broad content category (one of four) and 2) compared to the TSIA1, an increased emphasis is placed on reasoning skills throughout.

Built-in test aids. In addition to the wide array of on-screen tools also available in TSIA1 (e.g., accessibility tools[3] and calculator), TSIA2 introduces an optional highlighter feature, allowing students to mark parts of passages or questions during a test. This feature has also been added to the TSIA2 Study App so that students have an opportunity to familiarize themselves with this new functionality as they prepare for testing.

The TSIA2 also retains several features of TSIA1. The features of TSIA1 that continue to be key elements in TSIA2 are discussed in the following paragraphs.

Text complexity. Text complexity is a measure of passages' inherent reading challenge irrespective of question complexity or difficulty. Passages included in the TSIA2 ELAR Tests, like those used in TSIA1, cover a specified range of text complexity aligned to college and career readiness levels of reading. Passages range from "somewhat challenging" to "highly complex," with "complex" reflecting the college and career readiness threshold. See Appendix A: Text Complexity (Qualitative)—Reading and Writing for details on the qualitative text complexity rubric.

[3] These accessibility tools include Wizard, Read&Write Gold, the NonVisual Desktop Access (NVDA) Screen Reader, ZoomText® Magnifier/Reader, Kurzweil 3000, and JAWS.®

Words in context. On the TSIA2 ELAR Tests, test takers are called on to engage in close reading of texts and to derive the meaning of words and phrases from the contexts in which they are used. The skills and knowledge tested are broadly useful in numerous subject areas and careers. Some reading-focused questions assess vocabulary, including word- and phrase-meaning questions, in both extended contexts and single sentences. Test takers are presented with other vocabulary-related challenges as well by reading- and writing-focused questions, including questions requiring test takers to analyze word choice rhetorically and to improve the precision, concision, and context appropriateness of expression.

Command of evidence. In the ELAR Tests, test takers analyze material across a wide range of disciplines (humanities, social science, and science) and other contexts (practical affairs, human relationships, and career-related topics). They draw on textual evidence to support their answers and apply an understanding of how authors make use of evidence.

Standard English language conventions. Skilled expression in language requires an understanding of the conventions of standard written English—the developed ability to apply language conventions in the service not only of correctness of expression but also of varied rhetorical purposes. Many of the writing-focused questions on the ELAR Tests assess language conventions, while others address rhetorically effective language use in the context of multi-paragraph passages that test takers must revise and edit.

Disciplinary literacy. Students' literacy development should not be seen as merely the fostering of generic communication skills but rather as being grounded in making students familiar and skilled with the differing literacy demands of particular fields of study. Some of these differences involve vocabulary, text structures and features, the kinds of claims made, and the nature and sources of evidence used to support those claims. The range of texts included in the TSIA2 ELAR Tests supports the teaching and assessment of literacy skills across a wide range of disciplines.

Problems grounded in real-world contexts. Test takers engage with questions grounded in real world contexts and directly related to the work performed in college and career. The ELAR Tests include literature and literary nonfiction, but they also feature passages that students are likely to encounter in science, social science, and other majors and careers. The Mathematics Tests, which focus on applied reasoning skills essential for college and career readiness, feature multistep applications for solving problems in science, social science, career scenarios, and other real-life contexts.

Mathematics that matters most. TSIA2 is rooted in the philosophy of a deeper focus on fewer, more important topics in mathematics. In keeping with this philosophy, the Mathematics Tests focus on topics that have been identified as essential for college and career readiness. Applied reasoning questions are emphasized over questions disconnected from the mathematics curriculum. There is also a strong emphasis on both fluency with mathematical procedures and conceptual understanding.

For an overview of the tests available in TSIA2, see Table 1.1 in Section 1.3 of this manual. For a more in-depth look at the TSIA2 CRC and Diagnostic Tests and to view test specifications, see Section 3.2 of Chapter 3: Test Development Procedures. For details on TSIA2 score reporting, see Section 5.4 of Chapter 5: Interpretation and Application of Results. For details on the alignment of TSIA2 test content

to Texas standards, refer to Section 7.2: Content-Oriented Validity Evidence and Alignment in Chapter 7: Validity.

1.2 Background on Assessment

Brief History of Development

In 2012, College Board entered into a contract with the THECB to create a suite of assessments to help determine whether and at what level students are prepared to enroll and succeed in entry-level, credit-bearing college courses. TSIA1 was launched summer of the following year. This original suite of assessments contained several computer-adaptive tests and an essay test and was designed to assess the skills and knowledge of entering undergraduate students in the areas of reading, writing, and mathematics, with the ultimate goal of improving student success rates in Texas colleges. It offered three tests (Placement, DE Diagnostic, and ABE Diagnostic) in each of three areas: reading, writing, and mathematics. For more information on TSIA1, refer to the *Texas Success Initiative Assessment Technical Manual* (College Board, 2017).

In 2019, College Board was awarded a contract to create an updated version of TSIA1 to serve the same goal going forward. The updated version, TSIA2, measures students' readiness for college-level coursework in the general areas of English language arts and reading (ELAR) and mathematics.

The following is a brief timeline of the development of TSIA2:

Jul 2019 – Contract to deliver TSIA2 awarded to College Board

Jul/Aug 2019 – College Board, in consultation with the THECB, defined TSIA2 test purpose; reviewed and aligned proposed test content to Texas standards; established test design; and developed initial test specifications and standards alignment documents

Aug/Sep 2019 – Representatives from College Board and the THECB met remotely and reviewed proposed test specifications and alignments; College Board revised documents based on input received

Sep 2019 – Representatives from College Board met with THECB leadership and Texas faculty in Austin and reviewed updated specifications, alignments, and sets of sample questions that exemplify the knowledge and skills assessed; College Board made further revisions based on feedback

Oct/Dec 2019 – College Board assembled the TSIA2 question bank

Nov 2019 – Pretesting of new and revised questions began

Feb 2020 – Representatives from College Board met with THECB leadership and Texas faculty in Austin to review and discuss the assembled TSIA2 question pool

July 2020 – Virtual standard setting conducted for the TSIA2 CRC Tests; diagnostic levels closely aligned to the National Reporting System Educational Functioning Levels (NRS EFL)[4] for diagnostic tests set

Sep/Nov 2020 – Pretest data analyzed; question bank for launch finalized

Jan 2021 – TSIA2 launched statewide

ELAR and Mathematics Design Goals

The design of the TSIA2 ELAR Tests is intended to accomplish the following goals:

- Creating integrated classification and diagnostic testing so that test takers move seamlessly through the CRC and Diagnostic Tests in a single experience and so that test takers who are placed in the not college ready category do not end testing without receiving actionable feedback
- Combining the separate multiple-choice Reading and Writing components of the TSIA1 placement tests so that test takers taking the new TSIA2 ELAR CRC Test have a single, seamless testing experience[5]
- Shifting from the two classification scores and three scoring categories of TSIA1 (college ready, developmental education, adult basic education [ABE]) to a single college readiness classification score and two scoring categories (college ready, not college ready)
- Reducing the number of constraints on question selection relative to TSIA1, allowing the TSIA2 adaptive testing engine to perform more flexibly and efficiently
- Continuing to provide diagnostic test takers with actionable information about academic strengths and weaknesses across a range of content-based strands
- Integrating TSIA1's standalone ABE testing components into TSIA2 diagnostic testing, providing a sufficient span of question difficulty in the new Diagnostic Test to cover the range previously addressed by separate TSIA1 diagnostic and ABE testing (i.e., six National Reporting System Educational Functioning Levels [NRS EFL]) and addressing the performance expectations outlined in the Texas Adult Education and Literacy (AEL) Content Standards 2.0
- Calibrating the new Diagnostic Test questions to the same ability scale (theta) as the CRC Test questions
- Documenting, confirming, and, where necessary, improving alignment with current Texas academic and ABE literacy standards, specifically (1) Texas College and Career Readiness

[4] For the diagnostic strand proficiency descriptors, interim cut scores are set using an equipercentile linking approach, leveraging data from corresponding tests in TSIA1 and TSIA2. These cut scores are to be verified via a standard verification process in 2021.

[5] The separate multiple-choice Reading and Writing components of the TSIA1 DE Diagnostic and ABE Tests were similarly combined so that test takers taking the new TSIA2 ELAR Diagnostic Test have a single, seamless testing experience.

Standards (2018), (2) Texas Essential Knowledge and Skills (TEKS), English III (2017), (3) AEL Content Standards 2.0, and (4) NRS EFL

- Reducing the number of questions delivered in placement/CRC and diagnostic testing

Essay

The Essay Test, which remains unchanged from TSIA1, is a constructed-response test designed primarily to ascertain (in conjunction with the CRC Test and possibly the Diagnostic Test) whether test takers are college ready or not college ready with respect to writing.

The Importance of Test Practice and TSIA2 Success

The AERA/APA/NCME Standards for Educational and Psychological Testing, Standard 8.1, state:

> *Information about test content and purposes that is available to any test taker prior to testing should be available to all test takers. Shared information should be available free of charge and in accessible formats (AERA, APA, & NCME, 2014, p. 133).*

In keeping with this standard and College Board's belief in providing the same access to information and opportunities to all test takers, all TSIA2 test takers are given access to free study and review resources. These resources, designed to help students identify and fill in skill and knowledge gaps prior to testing, include the Study App, which contains sample and practice questions for multiple-choice tests written or reviewed by College Board and available in accessible formats. The Study App allows students to not only preview the design and format of a TSIA2 test but also to experience responding to questions on a computer and using tools available in the actual testing environment, such as the highlighter and calculator. Students taking the Essay Test have access to free guides that provide test information and two sample essays for each of the available score points along with annotations that explain why each sample essay was given the indicated score.

As mentioned previously, students also have access to TSIA2 Learning Resources, an online platform that features materials and activities designed to support repeated review and practice. These materials are available at no charge to students and can be used either prior to testing or with a personalized Learning Locator Code after diagnostic testing.

This ready access to free practice and review resources achieves two key goals. First, these resources give students who use them multiple opportunities to study at their own pace and demonstrate what they have learned and can do. More importantly, by ensuring that these materials are accessible to all Texas students, including low-income, underrepresented, and underserved students as well as nontraditional students, we uphold College Board's pledge to adhere to the standards for test takers' rights. Standard 8.0 in the *Standards for Educational and Psychological Testing* states:

> *Test takers have the right to adequate information to help them properly prepare for a test so that the test results accurately reflect their standing on the construct being assessed and lead to fair and accurate score interpretations (AERA, APA, & NCME, 2014, p. 133).*

For more information about free resources available to Texas students, visit https://accuplacer.collegeboard.org/students/prepare-for-accuplacer/tsia-texas-success-initiative-assessment.

1.3 Description of Content

TSIA2's testing domain definitions are based on the highest quality information and resources available about the essential requirements for college and career readiness and success as well as Texas's own curriculum and assessment standards. College Board staff worked with education experts from across Texas to examine the evidence and define the domain of skills and knowledge to be measured in accordance with each assessment's primary purpose and the claims associated with each assessment. College Board test development staff also prepared test and question/task specifications that represent the depth and breadth of the defined domains and help ensure the consistent development of assessments of the highest quality.

This section provides an overview of the content of the tests and their question and task formats, in keeping with standards and best practices (AERA, APA, & NCME, 2014). Table 1.1 presents an overview of the test format for the TSIA2 Suite. For a more in-depth look at test content and specifications, see Chapter 3: Test Development Procedures. For the psychometric properties of the tests, see Chapter 6: Psychometrics.

Table 1.1:
The TSIA2 Suite of Assessments

	Number of Questions		
Test	**Discrete**	**Set-Based**	**Total**
TSIA2 ELAR CRC Test	22	8	30
TSIA2 ELAR Diagnostic Test	24	24	48
Essay Test	1 essay	-	1
TSIA2 Mathematics CRC Test	20	-	20
TSIA2 Mathematics Diagnostic Test	48	-	48

ELAR Overview

The TSIA2 ELAR suite consists of:

- a single multiple-choice College Readiness Classification (CRC) Test, providing (in conjunction with the Essay Test) information regarding test takers' college readiness in reading and writing;

- a single multiple-choice Diagnostic Test, providing information regarding test takers' academic strengths and weaknesses in reading and writing; and
- a constructed-response Essay Test.

Test takers move seamlessly between the various tests in the suite based on the routing framework (see Scores, Routing, and Classifications, below). Test takers must complete all required testing before any information on their performance is yielded. This includes automatic routing to the Diagnostic Test for those who are placed in the not college ready category, so they do not end testing without receiving actionable feedback.

CRC Test. Like its predecessor, the ELAR CRC Test includes reading- and writing-focused elements; unlike its predecessor, the ELAR CRC Test is delivered to test takers as a cohesive testing experience instead of as separate reading and writing tests.

The TSIA2 ELAR CRC Test is designed primarily to ascertain (in conjunction with the Essay Test) whether test takers are college ready or not college ready with respect to reading and writing. The test consists of 30 questions and is intended to collect evidence in support of a broad claim about student performance:

> *Students can demonstrate college readiness proficiency in reading and writing.*

Reading passages on the test are literary as well as informational and cover a range of disciplines (e.g., literature, humanities, social science, and science) and other topics (e.g., practical affairs, and human relationships). Both single and paired passages are included. The reading-focused test pool includes both authentic texts (i.e., previously published passages excerpted or minimally adapted from their published form) and commissioned texts (i.e., written specifically for the test). Writing passages, informative/explanatory in text type, are commissioned and sampled from the same range of disciplines and topics as the reading passages.

Questions are multiple-choice in format and are discrete (i.e., standalone) or part of sets built around a common passage or passages. Questions assess four broad knowledge and skill categories, two reading-focused and two writing-focused:

Reading-Focused

- Literary text analysis
  - Explicit information
  - Inferences
  - Author's craft
  - Vocabulary

- Informational text analysis and synthesis
  - Main ideas and supporting details
  - Inferences (single-passage)
  - Author's craft
  - Vocabulary (interpreting words and phrases in context)
  - Synthesis (paired argumentative passages)

Writing-Focused

- Essay revision and editing
  - Development
  - Organization
  - Effective language use
  - Standard English conventions
- Sentence revision, editing, and completion
  - Conventions of grammar
  - Conventions of usage
  - Conventions of punctuation

Quick Facts:

- The computer-adaptive ELAR CRC Test has 30 questions; the linear, accommodated COMPANION form has 44 questions.
- All questions are multiple-choice.
- Questions may be discrete or set-based.
- One overall ELAR CRC Test score, ranging from 910 to 990, is reported.

For a more in-depth look at test content and specifications, see Chapter 3: Test Development Procedures.

Diagnostic Test. The ELAR Diagnostic Test is designed primarily to identify test takers' academic strengths and weaknesses with respect to reading and writing. The test consists of 48 questions.

Like the ELAR CRC Test, the Diagnostic Test is delivered to test takers as a cohesive experience that includes reading- and writing-focused questions. It subsumes the separate TSIA1 DE Diagnostic and ABE tests while encompassing the same range of question difficulty as the prior two tests.

As with the ELAR CRC Test, Diagnostic Test reading passages are literary as well as informational and cover a range of disciplines (literature, humanities, social science, science) and other topics (practical

affairs, human relationships). Both single and paired passages are included. The reading-focused test pool includes both authentic texts (previously published passages excerpted or minimally adapted from their published form) and commissioned texts (written specifically for the test). Writing passages, informative/explanatory in text type, are commissioned and sampled from a range of disciplines (humanities, social science, science) and other topics (practical affairs, human relationships).

Questions are multiple-choice in format and are discrete (standalone) or part of sets built around a common passage or passages. Similar to ELAR CRC Test questions, ELAR Diagnostic Test questions assess four broad knowledge and skills categories, two of which (reading-focused) compose the Text Analysis and Synthesis strand and two of which (writing-focused) comprise the Content Revision and Editing for Conventions strand.

Text Analysis and Synthesis Strand

- Literary text analysis (explicit information, inferences, author's craft, vocabulary)
- Informational text analysis and synthesis (main ideas and supporting details, inferences [single-passage], author's craft, vocabulary [interpreting words and phrases in context; decoding and recognizing words], synthesis [paired argumentative passages])

Content Revision and Editing for Conventions Strand

- Essay revision and editing (development, organization, effective language use, Standard English conventions)
- Sentence revision, editing, and completion (conventions of grammar; conventions of usage; conventions of punctuation; conventions of spelling and capitalization; purpose and organization; sentence combining)

Quick Facts:

- The computer-adaptive ELAR Diagnostic Test has 48 questions; the linear, accommodated COMPANION form has 72 questions.
- All questions are multiple-choice.
- Questions may be discrete or set-based.
- A diagnostic profile consisting of two elements is reported based on the test taker's performance on the Diagnostic Test: a diagnostic level closely aligned to the National Reporting System Educational Functioning Levels (NRS EFL) and proficiency descriptors with accompanying statements regarding the test taker's achievement on each content strand.

For a more in-depth look at test content and specifications, see Chapter 3: Test Development Procedures.

Essay. The TSIA2 Essay Test is a constructed-response test intended to collect evidence in support of a broad claim about student performance:

> *Students can demonstrate college readiness proficiency in writing.*

The Essay Test consists of one prompt, which includes a short passage and an assignment that states the writing task. In response to the prompt, students write an essay of 300 to 600 words. The test measures the extent to which test takers are able to consider a given topic, draw on their own ideas and experiences, and construct a multi-paragraph essay that states a position and supports its merit.

Test takers' essays, electronically scored on a holistic rubric by an automated essay scoring engine called the Intelligent Essay Assessor (IEA), are judged on six dimensions: purpose and focus; organization and structure; development and support; sentence variety and style; mechanical conventions; and critical thinking.

The score on this test, in conjunction with that on the multiple-choice ELAR CRC Test, helps colleges determine whether a student is ready for college-level coursework. Scores range from 1 to 8. An essay receives a 0 if it is too short to be evaluated, written on a topic other than the one presented, or written in a language other than English.

Like the multiple-choice tests, the Essay Test is typically computer delivered. Accommodated formats are also available for students with documented disabilities.

Quick Facts:

- The TSIA2 Essay Test is administered on the computer; two accommodated COMPANION forms are available.
- Essays are scored electronically.
- A holistic placement test score, ranging from 1 to 8, is reported, along with more detailed dimension scores and descriptors.

For more details on the TSIA2 Essay Test, see Chapter 3: Test Development Procedures.

ELAR Scores, Routing, and Testing Outcomes

The TSIA2 ELAR CRC and Diagnostic Tests have the following scores, routing paths, and classifications.

Scores

CRC Test. The multiple-choice ELAR CRC Test yields a score from 910 to 990. The test has a single college readiness classification score of 945, established through a standard setting process, and two scoring categories: college ready and not college ready.

Important: While CRC test takers' scores fall into either a college ready or not college ready range, test takers do not receive a college ready or not college ready designation based solely on CRC performance.

The final determination, as is made clear in the "Routing" and "Classifications" sections, below, is made in conjunction with performance data from the Diagnostic and/or Essay Tests. Test takers must complete all routed testing to receive any information regarding their performance.

Diagnostic Test. The multiple-choice ELAR Diagnostic Test yields the following information:

1. A classification into one of five diagnostic levels closely aligned to the NRS EFL:
    a. Level 2: Beginning Basic (subsumes Level 1: Beginning Literacy, for reporting purposes)
    b. Level 3: Low Intermediate
    c. Level 4: High Intermediate
    d. Level 5: Low Adult Secondary
    e. Level 6: High Adult Secondary

Important: Level 5 represents the college readiness cut score established in ELAR CRC standard setting. If a test taker's ELAR Diagnostic Test yields a diagnostic level of 5 or 6 and their performance on the Essay Test is at or above the college readiness classification score of 5, then the test taker's testing experience ends with a college ready classification. This represents test takers' second chance for receiving a college ready designation in ELAR.

2. A proficiency level (Basic, Proficient, or Advanced) that identifies the test taker's relative academic strengths and weaknesses in two content strands:
    a. Text Analysis and Synthesis (reading-focused)
    b. Content Revision and Editing for Conventions (writing-focused)

For each proficiency level, a proficiency statement describing expected performance at that level is available. Collectively, these statements allow test takers and/or their instructors to see what they know and can do in the given content category for each tier of performance (Basic, Proficient, or Advanced) and develop strategies for improvement. To view proficiency statements for TSIA2, see Appendix B: Proficiency Statements for TSIA2 Diagnostic Tests.

Essay Test. The Essay Test yields a single holistic score ranging from 1 to 8. The test has a single college readiness classification score, set at 5. In addition to the reported holistic score, feedback is provided on the six dimensions on which responses are evaluated, each of which is considered essential to a well-written essay: purpose and focus; organization and structure; development and support; sentence variety and style; mechanical conventions; and critical thinking.

Routing

Within Tests. Within the computer-delivered multiple-choice ELAR CRC and Diagnostic Tests, test takers are adaptively routed. (Nonstandard format accommodated versions of the tests are fixed-form linear tests.) The Essay Test is a single task and is therefore not adaptive.

Between Tests. The following section narrates the TSIA2 ELAR suite routing framework. The same information is represented visually in Figure 1.1: ELAR Routing Flow.

1. CRC Test

All test takers are administered the CRC Test first.

   a. If the CRC Test yields a score in the college ready range (i.e., a score at or above the college readiness classification score), then test takers are routed to the Essay Test.
   b. If the CRC Test yields a score in the not college ready range (i.e., a score below the college readiness classification score), then test takers are routed to the Diagnostic Test.

In the "b" scenario, a path to Essay remains should test takers attain a diagnostic level of 4 or higher; see below.

2. Diagnostic Test

Test takers are routed to the Diagnostic Test if the CRC Test yields a score in the not college ready range. Test takers then experience one of two scenarios:

   a. If performance on the Diagnostic Test yields a diagnostic level of 4 or higher, then test takers are routed to the Essay Test.
   b. If performance on the Diagnostic Test yields a diagnostic level of 3 or lower, then test takers have not demonstrated college readiness on the Diagnostic Test. These test takers receive an individual score report (ISR) indicating that they are not college ready.

3. Essay

Test takers may be routed to the Essay Test in one of two ways:

   a. If the CRC Test yields a score in the college ready range, then test takers are routed to the Essay Test.
   b. If the Diagnostic Test yields a diagnostic level of 4 or higher, then test takers are routed to the Essay Test.

Important: While all test takers receiving a diagnostic level of 4 or higher are routed to the Essay, only those test takers receiving a diagnostic level of 5 or 6 are eligible to receive a college ready designation.

Test takers' performance on the Essay Test results in one of five scenarios:

If test takers are routed to the Essay Test from the CRC Test:

A. If performance on the Essay Test is at or above the college readiness classification score, then test takers have demonstrated college readiness on both the CRC Test and the Essay Test. These test takers receive an ISR indicating that they have demonstrated college readiness.

B. If performance on the Essay Test is below the college readiness classification score, then test takers have not demonstrated college readiness on the Essay Test. These test takers receive an ISR indicating that they have not demonstrated college readiness; they may attempt to retake the Essay Test, or they may pursue placement in a corequisite course.

If test takers are routed to the Essay Test from the Diagnostic Test:

A. If the Diagnostic Test yields a diagnostic level of 5 or 6 and performance on the Essay Test is at or above the college readiness classification score, then test takers have demonstrated college readiness on both the Diagnostic Test and the Essay Test. These test takers receive an ISR indicating that they have demonstrated college readiness.
B. If the Diagnostic Test yields a diagnostic level of 5 or 6 but performance on the Essay Test is below the college readiness classification score, then test takers have not demonstrated college readiness on the Essay Test. These test takers receive an ISR indicating that they have not demonstrated college readiness; they may attempt to retake the Essay Test, or they may pursue placement in a corequisite course.
C. If the Diagnostic Test yields a diagnostic level of 4, then test takers have not demonstrated college readiness on the Diagnostic Test. Irrespective of their performance on the Essay Test, these test takers receive an ISR indicating that they have not demonstrated college readiness.

Testing Outcomes

Following testing, test takers who have completed all tests to which they are routed receive an individual score report (ISR) generally showing either a **college ready** classification or a **diagnostic** profile. Test takers bypassing the Diagnostic Test who score a 5 or below on the Essay Test receive an ISR that includes only their CRC Test score and Essay Test holistic score and dimension statements.

1. College ready classification

Test takers may receive a college ready classification in one of two ways:

   a. Test takers who score in the college ready range on the CRC Test and whose performance on the Essay Test is at or above the college readiness classification score are classified as college ready.
   b. Test takers who score in the not college ready range on the CRC Test but receive a diagnostic level of 5 or 6 on the Diagnostic Test and whose performance on the Essay Test is at or above the college readiness classification score are classified as college ready.

2. Diagnostic profile

Test takers may receive a diagnostic profile in one of three ways:

a. Test takers who score in the college ready range on the CRC Test but whose performance on the Essay Test is below the college readiness classification score receive a diagnostic profile as part of their ISR following diagnostic testing.
b. Test takers who score in the not college ready range on the CRC Test and receive a diagnostic level of 4 or lower on the Diagnostic Test are given a diagnostic profile. (This is so even for test takers who receive a diagnostic level of 4 and are routed to the Essay Test; they are ineligible to receive a college ready designation.)
c. Test takers who score in the not college ready range on the CRC Test and receive a diagnostic level of 5 or 6 on the Diagnostic Test but whose performance on the Essay Test is below the college readiness classification score receive a diagnostic profile as part of their ISR.

As noted in "Scores," above, the diagnostic profile includes a diagnostic level and two proficiency descriptors along with proficiency statements accompanying the descriptors.

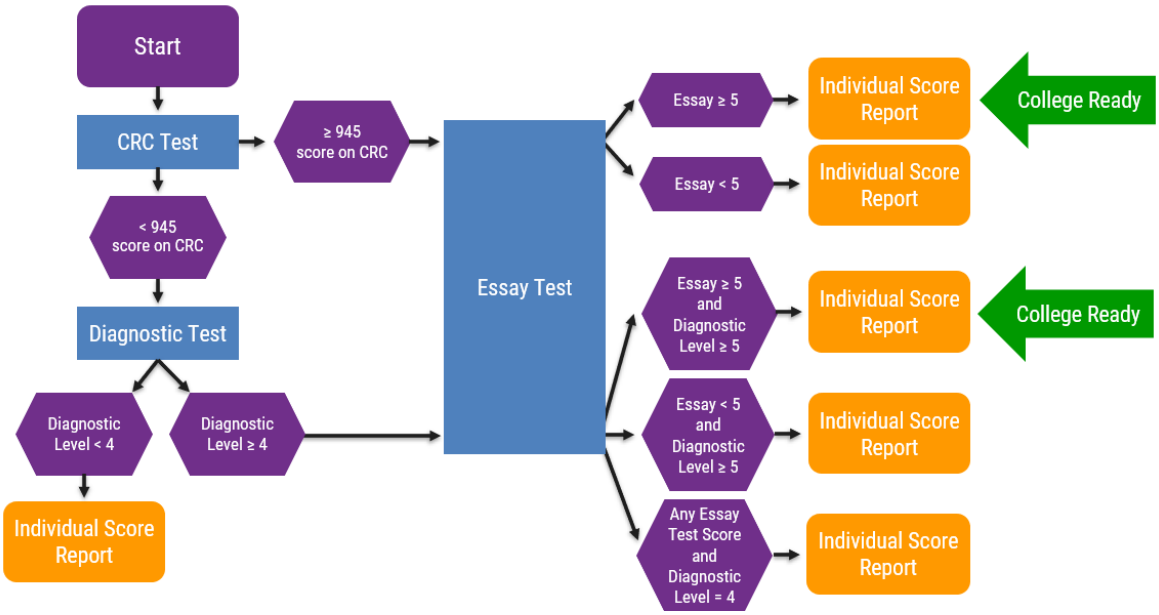

Figure 1.1: ELAR Routing Flow

Mathematics Overview

The TSIA2 Mathematics suite consists of

- a single multiple-choice College Readiness Classification (CRC) Test, providing information regarding test takers' college readiness in mathematics;
- a single multiple-choice Diagnostic Test, providing information regarding test takers' academic strengths and weaknesses in mathematics.

Test takers move seamlessly between the CRC and Diagnostic Tests based on the routing framework (see Scores, Routing, and Classifications, below). Test takers must complete all required testing before any information on their performance is yielded. This includes automatic routing to the Diagnostic Test for those who are placed in the not college ready category, so they don't end testing without receiving actionable feedback.

CRC Test. The Mathematics CRC Test is designed primarily to ascertain whether test takers are college ready or not college ready with respect to mathematics. The test consists of 20 questions and is intended to collect evidence in support of a broad claim about student performance:

> *Students can demonstrate college readiness proficiency in mathematics.*

Questions are multiple-choice in format and are discrete. Questions assess four broad knowledge and skill categories:

- Quantitative Reasoning
  - Calculating ratios, proportions, and percentages
  - Identifying, manipulating, and interpreting linear equations and expressions
- Algebraic Reasoning
  - Solving equations (linear, quadratic, polynomial, exponential, rational, and radical)
  - Evaluating functions
  - Solving algebraic problems in context
- Geometric and Spatial Reasoning
  - Converting units within measurement systems
  - Solving geometric problems (perimeter, area, surface area, and volume)
  - Performing transformations
  - Applying right triangle trigonometry
- Probabilistic and Statistical Reasoning
  - Classifying data
  - Constructing appropriate representations of data

- Computing and interpreting probability
- Describing measures of center and spread of data

Quick Facts:

- The computer-adaptive Mathematics CRC Test has 20 questions; the linear, accommodated COMPANION form has 30 questions.
- All questions are multiple-choice.
- All questions are discrete.
- Basic, square root, and graphing calculators are allowed on some questions on the computer-adaptive test[6]; a square root calculator is allowed on the COMPANION test.
- One overall Mathematics CRC Test score, ranging from 910 to 990, is reported.

For a more in-depth look at test content and specifications, see Chapter 3: Test Development Procedures.

Diagnostic Test. The Mathematics Diagnostic Test is designed primarily to identify test takers' academic strengths and weaknesses with respect to mathematics. The test consists of 48 questions.

As in the Mathematics CRC Test, all questions are multiple-choice in format and are discrete. Test questions cover the same four broad knowledge and skill categories as on the CRC Test but include additional content to evaluate students at lower-performing levels:

- Quantitative Reasoning
  - Calculating ratios, proportions, and percentages
  - Identifying, manipulating, and interpreting linear equations and expressions
- Algebraic Reasoning
  - Solving equations (linear, quadratic, polynomial, exponential, rational, and radical)
  - Evaluating functions
  - Solving algebraic problems in context
- Geometric and Spatial Reasoning
  - Converting units within measurement systems
  - Solving geometric problems (perimeter, area, surface area, and volume)
  - Performing transformations

[6] If a question is configured to allow for the use of a calculator, the dropdown calculator icon will present in the top right corner of the screen. For questions that are configured for multiple calculators, clicking on the icon will provide the student with a drop-down menu that could include two or three of the following: Basic (or four-function) calculator; square root calculator (or four-function calculator with a square root button); and TI-84 graphing calculator.

- Applying right triangle trigonometry

- Probabilistic and Statistical Reasoning
  - Classifying data
  - Constructing appropriate representations of data
  - Computing and interpreting probability
  - Describing measures of center and spread of data

Quick Facts:

- The computer-adaptive Mathematics Diagnostic Test has 48 questions; the linear, accommodated COMPANION form has 72 questions.
- All questions are multiple-choice.
- All questions are discrete.
- Basic, square root, and graphing calculators are allowed on some questions on the computer-adaptive test[7]; a square root calculator is allowed on the COMPANION test.
- A diagnostic profile consisting of two elements is reported based on the test taker's performance on the Diagnostic Test: a diagnostic level closely aligned to the National Reporting System Educational Functioning Levels (NRS EFL) and proficiency levels with accompanying statements regarding the test taker's achievement on each content strand.

For a more in-depth look at test content and specifications, see Chapter 3: Test Development Procedures.

Mathematics Scores, Routing, and Testing Outcomes

The TSIA2 Mathematics CRC and Diagnostic Tests have the following scores, scoring categories, and routing paths.

Scores

CRC Test. The multiple-choice Mathematics CRC Test yields a score from 910 to 990. The test has a single college readiness classification score of 950, established by a standard setting process, and two scoring categories: college ready and not college ready.

[7] If a question is configured to allow for the use of a calculator, the dropdown calculator icon will present in the top right corner of the screen. For questions that are configured for multiple calculators, clicking on the icon will provide the student with a drop-down menu that could include two or three of the following: Basic (or four-function) calculator; square root calculator (or four-function calculator with a square root button); and TI-84 graphing calculator.

Diagnostic Test. The multiple-choice Mathematics Diagnostic Test yields the following information:

1. A classification into one of five diagnostic levels closely aligned to the NRS EFL:
    a. Level 2 Beginning Basic (subsumes Level 1: Beginning Literacy, for reporting purposes)
    b. Level 3: Low Intermediate
    c. Level 4: Middle Intermediate
    d. Level 5: High Intermediate
    e. Level 6: Adult Secondary

Important: Level 6 represents the college readiness cut score established in Mathematics CRC standard setting. Test takers whose Mathematics Diagnostic Test yields a diagnostic level of 6 are deemed college ready in mathematics. This represents test takers' second chance for receiving a college ready designation in mathematics.

2. A proficiency descriptor (Basic, Proficient, or Advanced) that identifies the test taker's relative academic strengths and weaknesses in four content strands:
    a. Quantitative Reasoning
    b. Algebraic Reasoning
    c. Geometric and Spatial Reasoning
    d. Probabilistic and Statistical Reasoning

For each proficiency level, a proficiency statement describing expected performance at that level is available. Collectively, these statements allow test takers and/or their instructors to see what they know and can do in the given content category for each tier of performance (i.e., Basic, Proficient, or Advanced)." To view proficiency statements for TSIA2, see Appendix B: Proficiency Statements for TSIA2 Diagnostic Tests.

Routing

Within Tests. Within the computer-delivered multiple-choice Mathematics CRC and Diagnostic Tests, test takers are adaptively routed. (Nonstandard format accommodated versions of the tests are fixed-form linear tests.)

Between Tests. The following section narrates the TSIA2 Mathematics suite routing framework. The same information is represented visually in Figure 1.2: Mathematics Routing Flow.

1. CRC Test

All test takers are administered the CRC Test first.

    a. If the CRC Test yields a college ready score, mathematics testing is concluded.
    b. If the CRC Test yields a not college ready score (i.e., a score below the cut score), test takers are routed to the Diagnostic Test.

2. Diagnostic Test

Test takers are routed to the Diagnostic Test if their CRC Test yields a score in the not college ready range. Test takers then experience one of two scenarios:

a. If performance on the Diagnostic Test yields a diagnostic level of 6, then their testing experience ends with a college readiness classification.
b. If performance on the Diagnostic Test yields a diagnostic level of 5 or lower, then test takers have not demonstrated college readiness on the Diagnostic Test. These test takers receive an individual score report (ISR) indicating that they are not college ready.

Testing Outcomes

Following testing, test takers receive either a **college ready** classification or a **diagnostic** classification.

1. College ready classification

Test takers receive a college ready classification in one of two ways:

a. Test takers who score in the college ready range on the CRC Test are classified as college ready.
b. Test takers who score in the not college ready range on the CRC Test but receive a diagnostic level of 6 on the Diagnostic Test are classified as college ready.

2. Diagnostic profile

a. Test takers who score in the not college ready range on the CRC Test and receive a diagnostic level of 5 or lower based on the Diagnostic Test are given a diagnostic profile.

As noted in “Scores,” above, the diagnostic profile includes a diagnostic level and four proficiency descriptors along with proficiency statements accompanying the descriptors.

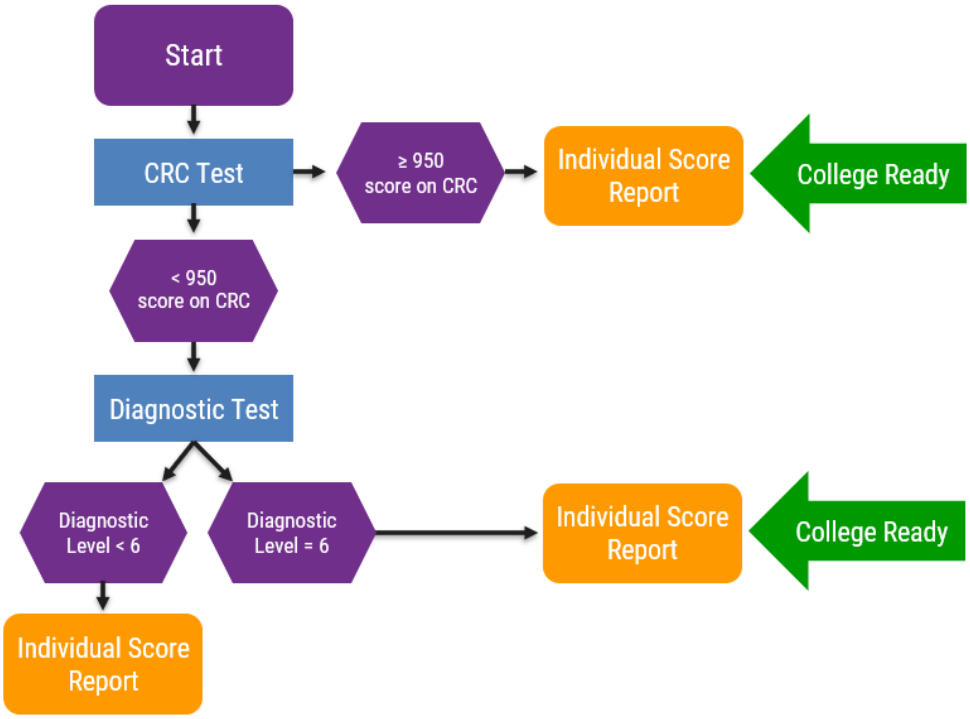

Figure 1.2: Mathematics Routing Flow

Chapter 2 — Fairness

Introduction

This chapter addresses the overarching issue of fairness. Section 2.1 covers College Board's general approach to fairness and explains the tenets to which all of our assessments must adhere in order to meet the AERA/APA/NCME Standards, industry best practices, and College Board's own internal standards. Section 2.2 shifts the focus to examine the content of the Texas Success Initiative Assessment 2.0 (TSIA2), in order to view fairness through the lens of construct-relevant content. Section 2.3 examines the steps that are taken by College Board at all stages of the "life of the test" in order to ensure fairness. This includes considerations related to test design, development, administration, scoring, and the interpretations of those scores. The section also looks at other processes put in place to ensure a fair testing experience, including matters of accessibility and accommodation.

2.1 Fairness in College Board Assessments

College Board believes in providing all test takers with a fair opportunity to demonstrate their achievement on the tests in the TSIA2 suite: the English Language Arts and Reading (ELAR) College Readiness Classification (CRC) and Diagnostic Tests, the Mathematics CRC and Diagnostic Tests, and the Essay Test (College Board, 2021a, 2021b). Conceptually, *fairness* can be defined in terms of both equitable treatment of all test takers in test administration and equal measurement quality across subgroups and populations. Best practices as well as standards 3.1–3.5 of the AERA/APA/NCME *Standards for Educational and Psychological Testing* call for test publishers to "minimize barriers to valid score interpretations for the widest possible range of individuals and relevant subgroups" (AERA, APA, & NCME, 2014, p. 63). An assessment should be built in such a way that the constructs being assessed are measured equitably for all intended test takers and test-taking subgroups; it should be administered in a manner that is fair and equitable for all test takers, regardless of gender, race/ethnicity, and other construct-irrelevant characteristics; and its results should be interpreted and used in ways that align with the intended purpose(s) of the assessment.

To accomplish these goals, four aspects of fairness, identified by the *Standards*, should be addressed when developing and administering an assessment.

1. **Fairness in treatment during the testing process.** Fairness in treatment involves "maximiz[ing], to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure" (AERA, APA, & NCME, 2014, 51). The *Standards* note that test makers have traditionally tried to meet this goal through standardization of the testing process—that is, by ensuring that all students are given the same instructions, testing time, and the like—but also that test makers increasingly recognize that "sometimes flexibility is needed to provide essentially equivalent opportunities for some test takers" (51) when accommodations (modifications) in testing don't compromise the construct being measured (e.g., reading achievement).

2. **Fairness as lack of measurement bias.** Per the *Standards*, bias in a measurement itself or in the predictions made from it may occur when "characteristics of the test itself that are not related to the construct being measured, or the manner in which the test is used, [lead to] different meanings for scores earned by members of different identifiable subgroups" (51). Bias in this sense can play out as differential performance on questions and/or tests by identified subgroups that are equally matched on the characteristic of interest (e.g., differential performance on a mathematics item by subgroups with equivalent mathematics achievement) and/or in differential predictions (inferences) about such subgroups. It is a responsibility of test makers to identify and root out such construct-irrelevant factors when they advantage or disadvantage subgroups of test takers.
3. **Fairness in access to the construct(s) being measured.** The *Standards* define accessible testing situations as those that "enable all test takers in the intended population, to the extent feasible, to show their status on the target construct(s) without being unduly advantaged or disadvantaged by individual characteristics (e.g., characteristics related to age, disability, race/ethnicity, gender, or language) that are irrelevant to the construct(s) the test is intended to measure" (52). Accommodations may take such forms as providing students with visual impairments access to large-print versions of text (when visual acuity isn't the construct being measured) and avoiding the use of regional expressions in test questions intended for a national or international audience.
4. **Fairness as validity of individual test score interpretations for the intended uses.** The *Standards* indicate that test makers and users should attend to differences among individuals when interpreting test data and not generalize about individuals from the performance of subgroups to which they belong. In practice, the *Standards* note, data on subgroup performance should not lead to the conclusion that subgroups are "homogeneous or that, consequently, all members of a group should be treated similarly when making interpretations of test scores for individuals (unless there is validity evidence to support such generalizations)" (53). Given those considerations, "adaptations to individual characteristics [e.g., reducing language barriers in testing when language proficiency isn't the construct being measured] and recognition of the heterogeneity within subgroups may be important to the validity of individual interpretations of test results in situations where the intent is to understand and respond to individual performance." At the same time, test makers also have to consider whether such adaptations may, for particular purposes, "be inappropriate because they change the construct being measured, compromise the comparability of scores or use of norms, and/or unfairly advantage some individuals" (53–54).

College Board embraces the fairness guidelines articulated by the *Standards* and the overarching goal of ensuring the maximal inclusiveness, representativeness, and accessibility of its test materials consistent with the constructs, purposes, and uses of the tests. Through its fairness-related documentation, processes, procedures, staff and consultant trainings, and other support materials, College Board strives to ensure that all its tests, including those in the TSIA2 suite:

- are appropriate for and accessible to a defined test-taking population as well as identified subgroups of that population;

- are appropriate as medium- to high-stakes assessments of college and career readiness;
- neither advantage nor disadvantage individual test takers or defined population subgroups of test takers due to factors not related to the constructs being measured (e.g., reading comprehension, mathematics achievement);
- are free of content or contexts likely to give offense, provoke a highly distracting emotional response, or otherwise inhibit test takers from performing their best on the tests;
- accurately and fairly portray the diverse peoples of the United States and the world and convey the widest possible range of ideas, perspectives, and experiences consistent with the tests' designs;
- are as fully and as widely accessible to as many test takers as possible through design and development processes yielding materials consistent with the principles of universal design and through a range of accommodations and supports for test takers with particular needs, while remaining faithful, to the fullest extent possible, to the constructs being measured; and
- have clearly articulated purposes for which they and their data should and should not be used.

This chapter provides a brief overview of several interrelated issues of fairness as they pertain to TSIA2—specifically, how College Board ensures the fairness of its test content, makes that content as accessible as possible, and provides accommodations and supports for students needing them. Preceding that discussion is a concise description of each test in the TSIA2 suite, which is intended to provide needed context for evaluating College Board's test fairness practices.

2.2 TSIA2 Constructs, Purposes, and Populations

As the AERA/APA/NCME *Standards* make clear at numerous points, test fairness cannot be evaluated separately from an understanding of what a given test is purporting to measure (its construct or constructs), what purpose or purposes it is intended to serve, and who comprises the test-taking population. Consideration of test construct is important because potential modifications to test content or delivery in the name of fairness should aim to eliminate or reduce artificial barriers to access while preserving, as much as possible, the essence of the knowledge, skills, and abilities (KSAs) being measured.

In other words, fairness in testing is, to a large extent, about minimizing construct-irrelevant factors precluding test takers from demonstrating what they know and can do. Providing a student who is visually impaired a test in large print or braille is very likely to be a reasonable modification when the construct being measured is reading (because while most people read visually, the underlying construct is comprehension of textual information) but not when the construct being measured is visual acuity. An understanding of test purpose is important as well because a test designed for one use—say, placement into a program—may or may not be suitable for another use—such as diagnosing deficiencies in performance—and because purpose informs the validity of inferences that can be drawn about test

takers from their performance on the test. It is also critical to understand the intended test-taking population, both in general and in terms of identified subgroups, so that test design and development can be guided to maximize accessibility for all test takers, test delivery can anticipate and accommodate the special needs of individuals in the population without compromising the construct(s) being measured, and test materials can be evaluated in relation to their suitability for the population and its constituent subgroups.

TSIA2 is the collective term for the suite of tests that includes two multiple-choice College Readiness Classification (CRC) Tests (one for ELAR and one for mathematics), two corresponding Diagnostic Tests, and an Essay Test. The CRC Tests are administered to entering college students to determine the degree to which they are prepared to succeed in college and workforce training programs. They are also administered to high school students who are eligible for dual enrollment either late in the junior year or early in the senior year in order to evaluate their college and career readiness. Students whose test performance does not meet the relevant college readiness classification benchmark are routed to the Diagnostic Tests, after which they may be placed in developmental or credit-bearing corequisite programs so that they may develop the skills and knowledge needed to succeed in college. Under certain performance conditions, students are routed to the Essay Test to produce a written response to an assigned prompt in order to demonstrate their writing achievement.

Each test is designed to collect evidence in support of a broad claim about student achievement.

- ELAR: Students can demonstrate college readiness proficiency in reading and writing.
- Mathematics: Students can demonstrate college readiness proficiency in mathematics.
- Essay: Students can demonstrate college readiness proficiency in writing.

The primary purpose of TSIA2 is to determine the degree to which students are ready to succeed in college and workforce training programs (the latter often being offered at two-year postsecondary institutions). All assessment content aligns with this purpose and is developed in accordance with 1) Texas curriculum and standards and 2) test designs grounded in the best available evidence about essential prerequisites for college and career readiness and success. Each test within the TSIA2 suite is designed to collect evidence from student performance in support of a broad claim about what students know and can do, and each claim is aligned to the primary purpose of assessing college and career readiness. Because TSIA2 assesses the content that matters most for college and career readiness and success, the resulting scores provide meaningful information about a student's likelihood of succeeding in postsecondary education. TSIA2 results should not, however, be used as the sole source of information for high-stakes decisions about students' academic achievement.

TSIA2 provides data that are used principally by postsecondary and high school educators as well as by students. In keeping with best practices and the requirements outlined in the *Standards*, TSIA2's intended uses and the key interpretations for its primary users are discussed in the following paragraphs, with a rationale presented for each use.

Assessing and monitoring students' college readiness. TSIA2 CRC Test scores (in the case of ELAR, in conjunction with Essay Test scores and detailed dimensional score descriptions) serve as meaningful indicators of students' readiness for college and career training, while Diagnostic Test performance yields statements that identify gaps in knowledge and skill areas so that subsequent intervention and targeted practice may be provided to students, helping them become more prepared for postsecondary-level courses. These data (CRC Test scores and Proficiency Statements[8] from the Diagnostic Tests) help states, districts, and schools monitor what proportion of their student body needs additional supports and what proportion is ready for or has a high likelihood of success in college-entry coursework.

Making college course placement decisions. The multiple-choice CRC and Diagnostic Tests are intended for use in higher education to provide a better understanding of students' level of preparedness for college-level work so that colleges may make more informed course placement policies and decisions. The CRC Tests, along with the Essay Test in the case of ELAR, inform users about the readiness of test takers for college-level courses, while Diagnostic Test results provide information to help identify appropriate interventions or corequisites. Such policies and decisions should be verified empirically, with appropriate adjustments made as necessary, in order to promote positive student and institutional outcomes. To this end, institutions can avail themselves of College Board's Admitted Class Evaluation Service (ACES), a free online validity study service that helps institutions to gather the predictive validity evidence needed to make or improve placement policies and decisions (see https://aces.collegeboard.org/ for more information). Colleges and universities using scores from College Board assessments, including TSIA2, in their admission or placement policies are encouraged to make use of ACES to verify the appropriateness of their policies, as doing so gives them the ability to make more informed decisions about confirming or refining these as appropriate to meet institutional needs.

It is important to emphasize that TSIA2 scores and score descriptors should be considered alongside other factors, such as students' high school grade point averages (HSGPA), in making course placement decisions. Moreover, in no case should the scores and descriptors be the basis for limiting students' advancement opportunities (e.g., decisions that restrict student access to challenging coursework or discourage aspirations of attaining higher education).

While all tests across the TSIA2 suite provide information about a student's readiness, these scores should not be used as a measure to rank or rate teachers, educational institutions, districts, or states, and users should exercise care when attempting to interpret test results for a purpose other than the intended purposes described here. College Board is not aware of any compelling validation evidence to support the use of any TSIA2 test, or other educational achievement measures, as the principal source of evidence for teacher or school leader evaluation. Only when properly used and subjected to several constraints can assessment data be used *in conjunction with other educational outcome measures* to make inferences about school quality and educational quality, including teaching and learning.

[8] Proficiency statements are presented in Appendix B: Proficiency Statements for TSIA2 Diagnostic Tests.

The populations taking TSIA2 as a group are inclusive of students in high schools and postsecondary institutions. These students are residents of Texas. In determining fairness policies and practices, College Board attends not only to these populations as wholes but also to identified subgroups within these populations. Consideration of such subgroups in test construction and administration is warranted by the AERA/APA/NCME *Standards* (2014) for a variety of reasons, including as part of efforts to eliminate measurement bias and to ensure equal access of all students and student subgroups to the test construct(s) being measured. In furtherance of these aims, College Board considers a number of population subgroups as part of qualitative and (when sample sizes permit) quantitative fairness analysis. These subgroups include (but are not necessarily limited to) male and female test takers and Black/African American, Asian American, Latinx, Native American, multiracial, and White test takers.

2.3 Fairness of TSIA2 Assessments

Test Construction

College Board has taken and continues to take numerous, exacting steps to establish and maintain the fairness of all its tests, including those in TSIA2. These efforts begin with the test design and continue through ongoing test development.

Test Design

Key Concepts. Fairness in test content starts with a thoughtfully crafted and sharply focused assessment design. With the exception of the Essay Test, which remains unchanged, all tests in TSIA2 were redesigned to ensure appropriate measurement of the KSAs that are critical for college and career readiness in Texas. The resultant TSIA2 suite is purpose-built using Texas's latest curriculum and standards as well as the best available evidence in order to measure attainment of essential college and career readiness and success prerequisites (in the case of the CRC Tests) and help identify the instructional supports needed to prepare students for successfully navigating postsecondary coursework (in the case of the Diagnostic Tests). A central tenet of the design philosophy was that the tests would address in depth the core knowledge and skills that evidence shows matter most. The content assessed includes close reading of texts; knowledge of words and phrases as they are used in context; command of evidence; writing skills required for presenting ideas logically, clearly, and correctly; mathematics concepts, skills, and practices strongly associated with the requirements of a wide range of college majors and workforce training programs; and problems grounded in real-world contexts. In addition, the design of assessment tasks aligns with best classroom teaching practices, thereby reducing the distance between assessment and instruction and making the test content more meaningful and accessible to students.

Evidence Gathering and Consultation. The design and development of TSIA2 mirrors the process College Board undertakes with all its assessments: intensive research, the convening of advisory committees, and solicitation of feedback from faculty members and subject matter experts representing a broad cross section of high school and higher education institutions. This prerelease process continues to be augmented by ongoing research as well as feedback from Texas high schools and institutions of

higher education. Some of the most prominent sources of evidence and feedback used in the design and development of TSIA2 are discussed in the following paragraphs.

Texas curriculum and standards. Just as the design and content of TSIA1 were guided by Texas curriculum and literacy standards, the same approach was taken in the creation of TSIA2. Specifically, test blueprints and subsequent content selection were based on (1) Texas College and Career Readiness Standards (2018), (2) Texas Essential Knowledge and Skills (TEKS), English III (2017), and Algebra II (2012), (3) Adult Education and Literacy (AEL) Content Standards 2.0, and (4) National Reporting System Educational Functioning Levels (NRS EFL) Descriptors. Aligning the tests in TSIA2 to these elements ensures the tests will continue to measure what constitutes core competencies in Texas and assess knowledge and skills that are considered to be essential for demonstrating college and career readiness.

Curriculum surveys. Approximately every three to four years, College Board undertakes and publishes the results from a survey of a nationally representative sample of middle school, high school, and postsecondary instructors. The primary purpose of the survey is to ascertain what skills and knowledge are deemed prerequisite for readiness for and success in common first-year, credit-bearing postsecondary courses across a range of subjects. The data act as a check on whether College Board college readiness assessments measure what postsecondary faculty deem critical for incoming students to know and be able to do. Based on these (and other) data, we make periodic refinements as warranted to our assessments in order to improve alignment to postsecondary faculty expectations and to better represent best instructional practices.

Test review committees. College Board convenes groups of secondary and postsecondary educators from around the country as independent consultants to review its test materials for both content soundness and fairness. Feedback from these groups is used to refine or remove material deemed problematic for use in its present form from the development process prior to its use with students in an operational setting. In addition to providing actionable information about specific test materials, these committees offer College Board developers ongoing, vital connections to and information about the teaching and learning that is undertaken in classrooms throughout the United States.

Academic advisory committees. When designing new assessments, College Board convenes academic advisory committees in the relevant subject areas. Composed of leading educators, these committees advise the organization on matters relating to educational philosophy, guiding principles, and standards for creating coherence among the instructional materials, assessments, and professional development programs and services designed to prepare students for college and workforce training success and on policy decisions regarding equity, parity, and access for all students.

In order to maximize the usefulness of the TSIA2 tests to Texas students and institutions, College Board test development staff met on two occasions with committees of Texas subject matter experts (assembled by the Texas Higher Education Coordinating Board (THECB)) to present proposed content and solicit input. In September 2019, test developers met with a committee to review test specifications, content alignments, and sets of sample questions that exemplify the knowledge and skills

assessed by TSIA2. College Board made revisions based on feedback from the committee, after which an initial question pool was assembled. In February 2020, the initial question pool was presented to a committee made up of a diverse group of content experts convened by the THECB. This committee provided feedback on the questions in the TSIA2 pool concerning the questions' appropriateness for assessing the core skills and knowledge assessed by each test, their suitability for the target population, and their alignment to Texas standards. The committee noted questions that succeed particularly well at assessing a given skill or element of knowledge, thereby offering College Board test developers positive models for future question development. The committee also flagged questions that might cause confusion or prompt other concerns, thereby offering models of question qualities to avoid in future development.

Content specifications. Content specifications operationalize the broad elements of a test's design in actionable, repeatable, and transparent detail. Content specifications describe such features as the subject matter and contexts to be included and the skills and knowledge to be measured. For TSIA2, content specifications were established after careful research and close consultation with the THECB and Texas subject matter experts and will be periodically reconsidered and refined as part of the evidence-gathering process described previously.

Among their virtues, detailed specifications for test materials—stimuli, questions, and the like—help further the goal of test fairness by ensuring that test content, regardless of when or by whom it was developed, meets the requirements of measuring the desired construct(s), aids in achieving the specified purpose(s) of the test, is suitable for the identified testing population and its subgroups, and is highly consistent in substance. In combination with other steps, such as item calibrations, the use of carefully articulated specifications helps ensure that test takers, regardless of the date on which they take the test or the particular set of test materials they receive, have highly comparable testing experiences and that their performance is not influenced by significant variability in test materials.

College Board has created and continues to maintain extensive documentation for use internally and by its partners in the development of stimulus materials (e.g., passages, informational graphics) and questions. These development guides include discussion of general issues, such as construct definition and identification of testing purposes and population as well as detailed process guidelines and examples of effective practice.

As feedback in various forms (e.g., student performance data, input from independent external reviewers, and comments from Texas subject matter experts and THECB officials) is received, refinements to content specifications and documentation for the purposes of improving the validity, reliability, transparency, and fairness of the assessments are occasionally made. When this happens, College Board communicates those changes to test takers and other stakeholders, including the THECB and users of TSIA2.

Test review guides. College Board maintains detailed content and fairness review guides for the independent experts it employs to evaluate its test materials. The content of these guides aligns closely

with that of the internal stimulus material and question development guides that direct College Board staff in their work. The review guides include information about the review process, expectations for reviewers, general content and fairness guidelines, and bullet lists of specific fairness considerations. These detailed guides promote fairness not only through explicit instructions about reviewing test content for fairness but also by virtue of standardizing and calibrating the review process, thereby helping ensure that different reviewers or groups of reviewers approach the review task in similar ways. These guides are periodically reviewed and updated to reflect any refinements to the test design. Updates may also be made to reflect changes in the guidance itself in response to changed circumstances and evolving understandings of fairness-related issues.

Test Development

Guided by detailed documentation and carefully defined processes undertaken by highly qualified academic subject matter and measurement experts, College Board's test development process for TSIA2 is designed to yield high-quality, valid, reliable, and fair assessments appropriate for the uses, populations, and subgroups identified earlier. As part of the development process, College Board staff employ various means, both qualitative and quantitative, to ascertain and maintain the fairness of test materials.

External Fairness Review. Prior to pretesting, all TSIA2 test materials are reviewed by external, independent reviewers who are asked to evaluate these materials for fairness. As a group, these reviewers are typically active classroom teachers drawn from across the nation, teach at the secondary and postsecondary levels, and are deeply familiar with both the student population of interest and its subgroups and the nature and purposes of the assessment. Reviewers are individuals from diverse backgrounds, live and work in different regions of the United States, and teach a range of disciplines, including English as a Second Language (ESL), at different levels (e.g., secondary, postsecondary), in different types (e.g., rural, suburban, urban) of institutions. Each review panel also includes diversity in gender representation. Reviewers are familiar with the test-taking population in general and with one or more population subgroups of interest. They also have expertise in and experience with best practices in diversity and inclusion. Reviews cover broad-based issues of fairness as well as specific matters with respect to race/ethnicity (including African American, Asian American, Latinx, American Indian/Native American, and mixed racial/ethnic backgrounds) and gender.

Fairness reviewers are charged with helping ensure that test stimuli and questions are broadly accessible to the wide-ranging student population that takes TSIA2 tests; do not advantage or disadvantage individual test takers or identified subgroups of test takers based on factors unrelated to the construct(s) being assessed; and address topics and texts that are appropriate for the audience of secondary and postsecondary students and the occasion of medium- to high-stakes testing. In addition to employing their own professional judgment and expertise, fairness reviewers are directed to apply criteria developed by College Board for each assessment program. These criteria include both general considerations and those specific to elements of individual tests (e.g., literature passages in ELAR). This fairness framework addresses the primary focus for qualitative fairness reviews; the concept of fairness;

the testing purposes and constructs; the audience for and occasion of testing; topic selection; individual and group portrayal; individual and group identification; ethnocentrism; and language. The framework's guidelines are sometimes expressly modified for particular kinds of content. For example, reviewers examining ELAR Test passages are informed that, in a limited exception to the general criteria on language, a small number of foreign words and phrases, slang terms, dialect, and/or idiomatic expressions may be acceptable in ELAR Test passages selected from works of U.S. and world literature, provided that sufficient context to enable understanding is available, because such elements are an authentic part of the real-world texts being sampled for the test.

While the criteria overviewed above are intended to be the primary basis for qualitative fairness evaluations of test materials, College Board encourages reviewers to draw on their professional judgment and expertise in order to apply the criteria flexibly and contextually. Moreover, reviewers are invited to raise issues that may not fall neatly into any of the above categories as part of the effort to ensure that all potential fairness issues receive thoughtful consideration.

Adhering to the practices implemented for all College Board tests, fairness reviews for all test materials are conducted before they are pretested with students. Reviewers first review the materials individually, followed by a discussion via meeting held remotely. Reviewers provide comments in advance of the meeting. These comments are read and considered by College Board staff, who prepare potential responses, such as edits or removal, for discussion at the meeting. College Board staff raise issues that were identified by reviewers as high priority, commented on by multiple reviewers, or not identified as high priority but nonetheless represent potentially serious matters. Guided by College Board staff, reviewers discuss these issues, evaluate College Board–proposed remedies when warranted, and raise issues of their own, either ones previously mentioned in advance comments or those newly discovered. College Board staff carefully assess all feedback, make decisions informed by best practices and expert consensus, and produce records of how particular issues were resolved. College Board staff have the latitude to make a range of revisions based on feedback (though less latitude to modify stimulus materials drawn from previously published sources) and may, if flaws are significant and/or pervasive, decide to stop further development and remove such questions from the pretest pool.

Pretesting. All questions are pretested on a motivated sample of test takers[9] that resembles the population of interest and is sufficient in size to allow College Board to evaluate the materials statistically in terms of difficulty, to discern whether the questions can differentiate between lower- and higher-achieving test takers, and to identify questions that test takers from different racial/ethnic and gender groups might have differentially responded to on the basis of construct-irrelevant factors. The data from at least 1,000 test takers responding to each question are used to evaluate the performance of the questions. Once questions have been pretested and the statistics associated with them have been computed, the materials are reviewed by measurement and content specialists for content soundness,

[9] Pretest questions are embedded into operational test administrations; test takers have no way of knowing which questions are scored and which are pretest questions (i.e., not scored).

fairness, statistical discrimination, difficulty, and differential performance among groups of tested students.

Differential Item Functioning (DIF) analysis. Analyses of differential item functioning, or DIF, are conducted on test questions at the pretest stage to identify any that may function differently for members of different population subgroups. It is important to note at the outset that DIF analysis is not based on the past test performance of various population subgroups, nor is DIF intended to remove questions from use that members of certain population subgroups do well on. Rather, DIF analysis serves to call attention to particular questions on which samples of certain population subgroups of equivalent achievement demonstrated a marked difference in performance.

The underlying assumption in conducting such analyses is that all test takers demonstrating the same level of achievement in the content area should have similar chances of answering each question correctly, regardless of subgroup membership. DIF occurs when individuals of different subgroups with similar achievement (i.e., similar scores on a test) differ notably in their performance on a specific test question. The presence of DIF provides a statistical indication that a question may function differently for individuals belonging to one subgroup than for those belonging to another subgroup who are at the same achievement level. Questions exhibiting DIF are divided into those showing low, medium, and high levels of DIF, with these designations based on established statistical thresholds. Those test questions exhibiting high levels of DIF have a greater-than-normal chance of measuring factors irrelevant to an assessment (such as those related to culture).

DIF analyses begin by examining any differences in performance on each individual question relative to two groups of comparable achievement, referred to as the *reference group* and the *focal group*. Questions identified after pretesting as exhibiting DIF over an established threshold, and thereby appearing to favor one group over another based on test taker samples matched on achievement, may undergo further review to determine whether some aspect of what the question is asking is drawing on one or more construct-irrelevant traits associated with subgroup membership (e.g., cultural background). As a result of DIF analyses, questions may be revised and repretested or eliminated from further use. For more information on DIF as it relates to TSIA2, see Chapter 3: Test Development Procedures.

Qualitative review is a critical complement to DIF analysis. The presence of DIF signals the possibility that a question may be biased, but the results of threshold DIF analyses alone do not determine whether a question is unfair. That judgment must be made by experts evaluating the question at hand, taking into consideration the purpose(s) of the assessment, the appropriateness of the question given the purpose(s) of the assessment (i.e., whether the knowledge or skill being tested falls within the test domain), and whether any construct-irrelevant factors are present in the question. Feedback from experts on questions flagged for DIF can be used to inform decisions about whether a given question should be excluded from the operational question pool and, more broadly, can help test developers identify or avoid introducing construct-irrelevant factors in future development.

The value of qualitative review is attested to by the *Standards for Educational and Psychological Measurement* (AERA, APA, & NCME, 2014). Although the *Standards'* discussions of DIF imply that DIF procedures are a common and expected part of test question analyses, the *Standards* also point out that statistical evidence of DIF (e.g., high DIF) does not necessarily imply a flaw or weakness in a question. In fact, the *Standards* suggest that when DIF occurs, test developers should try to "identify plausible explanations for the differences" (82) and then *may* choose to remove the question(s).

That said, the stronger the DIF violation, regardless of recognizable evidence of a flaw or weakness, the more consideration should be applied to the question's removal. Given that, TSIA2 policy with respect to DIF is that questions that exhibit low or medium DIF are retained for use unless internal and/or external content review identifies one or more construct-irrelevant factors likely contributing to the DIF results, while questions that exhibit high DIF are removed from further use until and unless they are revised and repretested (and again analyzed for DIF). A detailed description of DIF analysis for TSIA2 is in Section 3.3 Development of TSIA2 Assessments in Chapter 3: Test Development Procedures.

Operational Administration

Fairness also involves equality in test administration across all groups of test takers. For instance, detailed procedures are specified by College Board to ensure that each TSIA2 test is administered uniformly across all testing sites in a fair and equitable manner.[10] Without such standardization, the accuracy and comparability of score interpretations would be reduced (AERA, APA, & NCME, 2014).

TSIA2 tests may be administered in a variety of ways: at the institution where a student is enrolled or one that is convenient and authorized to deliver the test; with a human proctor or a virtual one; and as a computer-adaptive test or a linear,[11] accommodated (COMPANION) test. Additionally, COMPANION tests may be administered in one of several formats, including a paper-and-pencil test, braille, and audio CD, or a combination of these.[12] To ensure consistency in the administration of TSIA2 regardless of how or where a student is tested, College Board provides detailed guidelines and procedures for testing personnel, including:

- the Texas Success Initiative Assessment 2.0 Administrator's Manual (College Board, 2020a), which includes descriptions of tests and test purposes; cut scores and how they are set; appropriate use of the tests; test center and personnel responsibilities; secure handling of sensitive materials; testing accommodations; and administration of COMPANION tests.
- Reader scripts for each COMPANION test, which ensure all linear tests are read in the same way to all test takers, whether the test is delivered by a human reader or in a recorded format.

[10] Such uniformity, however, does not exclude the provision of appropriate accommodations for test takers with particular needs. See materials covering *Accessibility* and *Accommodations* in this chapter.

[11] Linear testing delivers all questions in the same order to each test taker, regardless of how earlier questions are answered.

[12] Depending on documented needs, test takers' accommodations may allow for the use of a braille test in combination with a CD.

Test security measures are also set in place to ensure that no test taker or group of test takers obtains access to information or opportunities that allow them to attain scores by fraudulent means and thereby jeopardize the validity and fairness of the results of the assessment. In addition to the security measures described in detail in Chapter 4 (Section 4.3: Security) of this manual, testing personnel must be certified to administer tests through the ACCUPLACER Certificate of Test Administration (ACTA) program, a mandatory assessment that covers topics such as testing policies, security protocols, and features of the testing platform. Testing personnel maintain certification by passing the ACTA test every year.

Predictive Validity Analyses

Fairness extends beyond question performance and test construction and is strongly tied to validity. Standard 3.7 (AERA, APA, & NCME, 2014) addresses the notion that fair assessments ensure validity of test score interpretations as a basis for predicting future performance. In the case of TSIA2, test scores should not provide different criterion prediction for different subgroups. A predictive placement validity study of the TSIA2 CRC Tests is planned for when sufficient data have been collected. College Board will perform exploratory investigation of differential prediction should it be warranted by placement results.

Score Reporting and Interpretation

A critical aspect of fairness is the fair and valid interpretations of test scores for intended uses (AERA, APA, &NCME 2014, p. 53). To support appropriate interpretations and inferences made on the basis of TSIA2 test scores, College Board provides detailed score reports to test takers (in the form of individual score reports, or ISRs) and test data users (in the form of institutional roster reports). Both types of reports are described in this section.

Along with publicly available interpretative materials developed by College Board, including content specifications, proficiency statements, dimension descriptions for the Essay Test, student brochures (described in this chapter), and guidance regarding intended uses of TSIA2 and TSIA2 test scores (presented in Chapter 1: Overview), score reports provide crucial information on test takers' preparedness for postsecondary work. Collectively, these materials are designed to make sure all test takers have access to the information they require so they may seek any appropriate just-in-time academic supports they need to connect them to success in college and career training programs and to help instructional programs in efforts to make such supports available. By providing such critical information to all test takers and the institutions and programs that serve them, College Board makes available to every test taker an equal opportunity to achieve success in college and workforce training programs.

Score Reporting

Test taker and institutional level reports have been developed for TSIA2. These reports are described in Chapter 5: Interpretation and Application of Results. Materials have also been developed to provide guidelines outlining proper interpretations and appropriate use of test results.

Score Interpretation and Proficiency Statements

Interpretation of TSIA2 scores conforms with the specific purposes of the tests. For the CRC Tests, test takers who score at and beyond the college readiness benchmarks are deemed college ready as described in the Texas College and Career Readiness Standards (CCRS). That is, a test taker who scores 945 or higher in the multiple-choice ELAR Test and 5 or higher in the Essay Test are deemed as having the KSAs necessary to succeed in entry-level community college and university English Language Arts courses. Similarly, a test taker who scores 950 or higher in Mathematics is deemed to have the KSAs necessary to succeed in entry-level community college and university mathematics courses. Test takers who score below the college readiness benchmark in either ELAR or Mathematics CRC Tests are routed to the corresponding Diagnostic Test. Results of the Diagnostic Tests are reported in two ways – Diagnostic Levels and Proficiency Levels, as described below.

Diagnostic Levels

Scores on the Diagnostic Tests are classified into and reported in terms of diagnostic levels 2 through 6. These levels are consistent with the NRS Educational Functioning Levels:

- ELAR
  - Level 2 – Beginning Basic (subsumes Level 1: Beginning Literacy, for reporting purposes)
  - Level 3 – Low Intermediate
  - Level 4 – High Intermediate
  - Level 5 – Low Adult Secondary
  - Level 6 – High Adult Secondary
- Mathematics
  - Level 2 – Beginning Basic (subsumes Level 1: Beginning Literacy, for reporting purposes)
  - Level 3 – Low Intermediate
  - Level 4 – Middle Intermediate
  - Level 5 – High Intermediate
  - Level 6 – Adult Secondary

The knowledge, skills, and abilities defining typical test taker performance at each level, appear in Appendix H of the Standard Setting Report (Bay and Duffy, 2020). Furthermore, scoring at Level 5 or higher in the ELAR Diagnostic Test is equivalent to meeting the college readiness benchmark for the multiple-choice ELAR CRC Test described above. Similarly, scoring at Level 6 in the Mathematics Diagnostic Test classifies the test taker as college ready in mathematics.

Proficiency Statements

In addition to diagnostic levels, test takers who take the Diagnostic Tests also receive a proficiency classification (Basic, Proficient, or Advanced) that identifies their KSAs in each diagnostic strand. Each

proficiency level is accompanied by statements describing what a typical test taker knows and can do in a given strand. These data-driven statements are intended to facilitate interpretation of performance on ELAR and Mathematics Diagnostic Test strands:

- ELAR
  - Text Analysis and Synthesis (Reading-focused)
  - Content Revisions and Editing for Conventions (Writing-focused)
- Mathematics
  - Quantitative Reasoning
  - Algebraic Reasoning
  - Geometric and Spatial Reasoning
  - Probabilistic and Statistical Reasoning

The proficiency statements offer useful information for understanding a test taker's level of attainment in each content strand. Furthermore, these statements allow educators and students to see the skills typically mastered at each score band so they may develop appropriate strategies for improvement.

Proficiency statements included in the ISRs are found in Appendix B: Proficiency Statements for TSIA2 Diagnostic Tests. For information on the development of the proficiency statements, see Chapter 5: Interpretation and Application of Results.

Essay Score and Dimension Descriptions

The TSIA2 Essay Test results include descriptions of holistic scores received by test takers. Each essay is evaluated based on its overall effectiveness, not on the basis of the individual writing characteristics in isolation.

The test taker also receives more detailed writing dimension scores and descriptions in the ISR. The detailed descriptions allow both educators and test takers to see the writing skills typically observed in essays at each dimension score point so that, as necessary, they may develop appropriate strategies for improvement. For an in-depth look at these descriptions, see Appendix C: TSIA2 Essay Scoring Rubrics.

Question Challenge Process

TSIA2's question challenge process offers additional transparency and a check on the soundness and fairness of test materials. Test takers may alert proctors to potential issues with test materials. These issues are then forwarded to College Board through established email channels. Such queries are routed to senior test development staff, who review the materials in question and, if applicable, develop answer explanations. If the question is from a study or practice resource, the reporting test administrator is supplied with the review's findings; if the question is still being used on an operational test, the reporting test administrator is advised that the material was reviewed and what the outcome of the process was. In the rare circumstance in which a review identifies a problem with the materials,

College Board undertakes appropriate remediating action, including removing problematic material from a question pool.

Practice

A critical aspect of test fairness rooted in but extending beyond design, development, and administration is practice. "Practice" includes the vital area of test familiarization—that is, making test takers aware of and comfortable with test instructions, formats, delivery methods, and the like. Resources and activities also focus on the underlying knowledge and skills fundamental to test constructs and hew closely to the test itself, their main purpose being to prepare test takers for the material they will encounter on test day.

Ensuring that all test takers have access to accurate, thorough information about the test well in advance of test day helps foster the goal of equity by giving everyone an equal chance to learn what is expected of them on the assessment, to address skill and knowledge gaps well in advance of the assessment, and to avoid spending valuable test time reading directions, figuring out what the questions are asking, and trying to understand how to navigate computer-delivered tests.

College Board provides a wealth of informational and practice-related materials for tests in the TSIA2 suite, all of them free of charge, to students and other stakeholders. These include full test specifications for the TSIA2 tests (College Board, 2021a, 2021b); the *Texas Success Initiative Assessment 2.0 Administrator's Manual* (College Board, 2020a); the *Texas Success Initiative Assessment 2.0 Technical Manual* (College Board, 2020b); sample test questions, along with answer explanations, available in the form of downloadable PDFs and via a Study App on the testing platform; Proficiency Statements designed to be used for interpreting Diagnostic Test scores and for guiding intervention prior to retesting; and two brochures for students, one intended for use before testing (*Texas Success Initiative Assessment 2.0 Student Informational Brochure*; College Board, 2020c) and one for use after testing (*Texas Success Initiative Assessment 2.0 Interpreting Your Scores*; College Board, 2020d)[13]. Additionally, College Board offers free training and professional development webinars designed to support faculty and test administrators who work with TSIA2 test takers. For more information on the practice materials and information provided, see Section 1.2 of this manual.

Accessibility

College Board is strongly committed to the concept of making TSIA2 test materials maximally accessible to all test takers. The organization subscribes to the principles of universal design, which, as noted by the AERA/APA/NCME *Standards*, has as its goal "[developing] tests that are as usable as possible for all test takers in the intended population, regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability" (AERA, APA, & NCME, 2014, p. 57). To make sure the largest number of test takers can access the tests, all TSIA2 materials are created in a highly

[13] Texas Success Initiative Assessment 2.0 *Interpreting Your Scores* is also included as a link in test takers' Individual Score Reports.

legible layout. Additional steps to provide maximum accessibility are discussed in the following paragraphs.

Ensuring the availability of assistive technology for online testing. Accessibility Wizard, which makes it possible for test takers with visual impairments to change the appearance of testing screens for easier viewing, is available in the online test environment. It gives test takers the ability to select a display that enhances the legibility of test materials presented, including the ability to choose a high-contrast color scheme, font and cursor/point size and color, and line spacing. Institutions that use other assistive technology as a standard accommodation for students whose vision impairments prevent them from accessing screen content or navigating with a mouse may elect to use these for administering TSIA2. Apart from providing the Accessibility Wizard, the testing platform can also leverage assistive technology, including Read&Write Gold[14], the NonVisual Desktop Access (NVDA) Screen Reader[15], ZoomText® Magnifier/Reader[16], Kurzweil 3000[17], and Job Access With Speech (JAWS®)[18].

Ensuring the availability of accommodated linear tests. All tests in the TSIA2 suite have two corresponding, comparable COMPANION forms. These are linear tests that present TSIA2 content in alternate formats. Designed for test takers who are not able to take computer-adaptive tests or for institutions that may be unable to administer them, COMPANION tests are available in several linear formats, including regular and large print "print-on-demand" test forms that test administrators may download from the platform, reader scripts, audio CDs, and braille. COMPANION forms use the same score scale as the computer-adaptive tests and are one-and-a-half times the length of their corresponding computer-adaptive tests.[19]

Ensuring ADA compliance. All print materials, including COMPANION tests, TSIA2 guides and manuals, sample questions, and test specifications are ADA compliant and can be used either in print forms or, if viewed on the computer, accessed through screen reader software.

Accommodations

While observing the principles and adopting the practices of universal design and of accessibility more generally are helpful in reducing the number and severity of construct-irrelevant barriers for all test takers, some test takers may still need additional support in order to complete the assessment and/or obtain valid test scores. Standards 3.9 through 3.14 discuss the responsibility of test makers to develop

[14] https://www.texthelp.com/en-us/products/read-write/

[15] http://www.nvaccess.org/

[16] http://www.aisquared.com/Products/index.cfm

[17] http://www.kurzweiledu.com/kurz3000.aspx

[18] http://www.freedomscientific.com/Products/Blindness/Jaws

[19] The computer-adaptive tests are shorter than their COMPANION counterparts because the adaptivity of the test engine creates greater testing efficiency by targeting materials to students' demonstrated achievement levels.

and provide test accommodations as well as the appropriate use of said accommodations (AERA, APA, & NCME, 2014, p. 67–70).

To provide a fair testing environment for all test takers, students with disabilities that affect their ability to participate in TSIA2 are eligible to test with the accommodations they need. While College Board takes steps to ensure TSIA2 is accessible to all test takers, testing programs at individual institutions, acting on the counsel of their Services for Students with Disabilities (SSD) coordinators and taking students' documented needs into consideration, are charged with determining and offering test takers accommodations for testing, such as large print, braille, or audio recording or a human reader, signer, or scribe. These testing programs, with support from College Board as needed, are entrusted with the responsibility of conducting testing without changing the construct or constructs being measured, such that scores maintain their meaning across all subgroups as well as for both accommodated and nonaccommodated test takers. This practice ensures that, when appropriate and possible, construct-irrelevant barriers that can interfere with test takers accurately demonstrating their true standing on a construct are removed (AERA, APA, & NCME, 2014). As previously discussed, a construct-irrelevant barrier is any factor unrelated to the concepts or characteristics the assessment is designed to measure that can lead to an unfair testing experience and distort test takers' scores, decreasing the validity of the scores for their intended uses.

In keeping with the AERA/APA/NCME *Standards* and best practices, accommodations are intended to "respond to specific individual characteristics, but [do] so in a way that does not change the construct the test is measuring or the meaning of scores" (AERA, APA, & NCME 2014, p. 67). To this end, all accommodated test formats and testing conditions are designed to be comparable, in that even though forms or conditions might be modified based on the needs of a particular test taker, the construct being tested and the meaning of the score remain unchanged.

The following are examples of accommodations intended to ensure eligible students receive the support they need. Please note that this list is not exhaustive.

Presentation. COMPANION tests are available for test takers who are not able to take computer-adaptive TSIA2 tests. These linear tests are offered in the following formats:

- Assistive technology compatible (i.e., screen reader–accessible) format
- Regular print
- Large print
- Braille
- Prerecorded audio (via CD)
- Human reader and reader script (Note: Reader reads entire test.)

Setting. Many accommodations are administered in the standard testing room. When judged appropriate, however, a Services for Students with Disabilities (SSD) coordinator may administer accommodated tests in nonstandard settings, including:

- Small-group setting
- One-to-one testing
- Private room
- Alternative testing site (with proctor present, either in person or virtually)

Additionally, test takers may be given preferential seating.

Responding. Test takers can also receive accommodations in how they record responses, including:

- Writer/Scribe
- Record answers on answer sheet

Timing and Scheduling. Finally, test takers can receive accommodations in timing and scheduling, including:

- Frequent breaks
- Specified time of day

Guidelines for Granting Accommodations

In general, students approved by their institution's Disability Support Services Office (DSSO) or similar office to receive testing accommodations meet the following criteria:

The student has a documented disability. Examples of disabilities include, but are not limited to, visual impairments, learning disorders, and medical impairments. A student must have documentation of their disability, such as a current psychoeducational evaluation or a report from a doctor. The type of documentation needed depends on the student's disability and the accommodations being requested.

Participation in an assessment is impacted. The disability must result in a relevant functional limitation that impacts the student's ability to participate in a TSIA2 test. For example, a student whose disabilities preclude sitting for extended periods may need accommodations, given that typically several TSIA2 tests are administered consecutively in a single sitting.

The requested accommodation is needed. The student must demonstrate the need for the specific accommodation requested. For example, a student requesting to be tested in a private room should have documentation showing that they have difficulty performing test tasks in an open setting.

Chapter 3 — Test Development Procedures

Introduction

This chapter discusses the design, development, and analysis involved in the creation of the Texas Success Initiative Assessment 2.0 (TSIA2). Section 3.1 defines the guiding principles behind the current test design. Section 3.2 provides the test specifications for English language arts and reading (ELAR), Essay, and Mathematics, for both the College Readiness Classification (CRC) and Diagnostic tests. Section 3.3 details the processes involved in the creation of test questions and the pretesting analysis that is undertaken (including analyses of difficulty, discrimination, and Differential Item Functioning [DIF]) for inclusion in the question pool. Section 3.4 provides an in-depth look at the Computer-Adaptive Test (CAT) algorithm and its features.

3.1 Guiding Principles of College Board's Test Development Process

College Board's test development process is guided by a set of principles, the consistent application of which helps ensure that every question and task that is ultimately selected for inclusion in a test is:

- evidence based and focused on the core set of knowledge and skills that are most important to prepare students for the rigors of college and career;
- measuring student skills and knowledge as directly and authentically as possible by employing a range of question and task types relevant to instruction and life;
- worth doing, crafted out of rich and engaging passages and contexts, reflective of best instructional practices, and rewarding of the academic excellence that any student can attain through deliberate practice;
- as motivating, interesting, engaging, and relevant to students as possible;
- written by experts, many of whom have teaching experience at the middle school, high school, and postsecondary levels;
- reviewed by multiple independent experts active in the field of education for content and fairness issues prior to pretesting;
- accessible and fair to all students, having been developed to be content relevant, accurate, authentic, and respectful in representation, and consistent with universal design principles.

In the development of TSIA2, College Board developers have also been also guided by current Texas academic and literacy standards, specifically (1) Texas College and Career Readiness Standards (CCRS) (2018), (2) Texas Essential Knowledge and Skills (TEKS), English III (2017) and Algebra II (2012), (3) Adult Education and Literacy (AEL) Content Standards 2.0, and (4) National Reporting System (NRS) Educational Functioning Levels (EFL).

Strict adherence to the above principles and standards helps ensure that TSIA2 deeply reflects the work that Texas students need to do to be ready for and to succeed in college and their career paths.

3.2 Test Specifications

This section presents detailed descriptions of each test in the TSIA2 Suite.

ELAR CRC Test

The ELAR CRC Test is designed primarily to classify (in conjunction with the Essay Test) test takers into college ready or not college ready categories with respect to reading and writing. The test consists of 30 questions and is intended to collect evidence in support of a broad claim about student performance:

Students can demonstrate college readiness proficiency in reading and writing.

In its standard form, the CRC Test is delivered adaptively via computer. A range of accommodated versions are available for test takers with documented disabilities that may prevent them from taking the computer-delivered assessments; tests in these formats are fixed-form linear (i.e., not adaptive).

Question format. All CRC questions are multiple-choice and represent a mixture of set-based and discrete questions.

Stimulus content. Reading-focused test stimuli include both authentic (i.e., previously published) passages and commissioned passages written for the test; literary passages as well as informational passages across a range of disciplines (i.e., literature, humanities, social science, and science) and other topics (e.g., practical affairs and human relationships); and single and paired passages. Passages are mostly informative/explanatory in text type, with some narratives and arguments; represent a range of text complexity centered on late secondary and early postsecondary bands; and range in length from 40 to 400 standard (i.e., six-character) words, with one passage in the range of 350 to 500 standard words.

Writing-focused test stimuli are commissioned passages sampled from a range of disciplines (i.e., humanities, social science, and science) and other topics (e.g., practical affairs and human relationships). Passages are primarily informative/explanatory in text type; represent a range of text complexities centered on late secondary and early postsecondary bands; and range in length from single sentences to prose passages of up to 350 standard words.

The range of text complexity test takers encounter in stimulus content is not distinctly constrained for. Instead, that range—which extends up to and includes college readiness and early postsecondary levels—is determined by associated question content.

Question content. Test questions assess four main categories, two reading-focused and two writing-focused.

Reading-Focused:

- Literary Text Analysis (i.e., explicit information, inferences, author's craft, and vocabulary)
- Informational Text Analysis and Synthesis (i.e., main ideas and supporting details, inferences [single-passage], author's craft, vocabulary [interpreting words and phrases in context], and synthesis [paired argumentative passages])

Writing-Focused:

- Essay Revision and Editing (i.e., development, organization, effective language use, and standard English conventions)
- Sentence Revision, Editing, and Completion (i.e., conventions of grammar, conventions of usage, and conventions of punctuation)

A single testing experience consists of 30 questions, half of which are reading-focused and half of which are writing-focused. Questions are presented in a seamless experience, with no break or division between reading-focused and writing-focused CRC questions. Reading-focused questions appear first, beginning with a Literary Text Analysis set. Writing-focused questions follow, beginning with an Essay Revision and Editing set.

The reading-focused element of the ELAR CRC Test consists of 15 questions:

- 1 4-question Literary Text Analysis set
- 11 discrete Informational Text Analysis and Synthesis questions
  - 2 Synthesis (paired argumentative passages) questions
  - 9 questions algorithmically chosen among all remaining Informational Text Analysis varieties

Similarly, the writing-focused element contains 15 questions:

- 1 4-question Essay Revision and Editing set
- 11 discrete Sentence Revision, Editing, and Completion questions algorithmically chosen among all Sentence Revision, Editing, and Completion varieties

Test Summary. Tables 3.1 and 3.2, below, present synopses of key aspects of the ELAR CRC and Diagnostic Tests.

ELAR Diagnostic Test

The ELAR Diagnostic Test is designed primarily to identify test takers' academic strengths and weaknesses with respect to reading and writing. In its standard form, the Diagnostic Test is delivered adaptively via computer. A range of accommodated versions are available for test takers with documented disabilities that may prevent them from taking the computer-delivered assessments; tests in these formats are fixed-form linear (i.e., not adaptive).

Question format. All diagnostic questions are multiple-choice and represent a mixture of discrete and set-based questions.

Stimulus content. Reading-focused test stimuli include both authentic (i.e., previously published) passages and commissioned passages written for the test. Diagnostic Test reading-focused passages are literary as well as informational; sample from a range of disciplines (i.e., literature, humanities, social science, science) and other topics (e.g., practical affairs and human relationships); and are single and paired. Passages are mostly informative/explanatory in text type, with some narratives and occasional arguments; represent a range of text complexity centered on late secondary and early postsecondary bands; and range in length from 40 to 400 standard (i.e., six-character) words.

Writing-focused test stimuli are commissioned passages sampled from a range of disciplines (i.e., humanities, social science, and science) and other topics (e.g., practical affairs and human relationships). Passages are primarily informative/explanatory in text type; represent a range of text complexity centered on late secondary and early postsecondary bands; and range in length from single sentences to prose passages of up to 350 standard words.

The range of text complexity test takers encounter in stimulus content is not distinctly constrained for. Instead, that range is determined by associated question content.

Question content. Test questions cover four main categories, two reading-focused and two writing-focused.

As with the ELAR CRC Test, the ELAR Diagnostic Test consists of questions in four content categories: two reading-focused (which together constitute the Text Analysis and Synthesis strand) and two writing-focused (which together constitute the Content Revision and Editing for Conventions strand). An asterisk (*) below denotes content on the Diagnostic Test not included on the CRC Test.

Text Analysis and Synthesis Strand

- Literary Text Analysis (i.e., explicit information, inferences, author's craft, and vocabulary)
- Informational Text Analysis and Synthesis (i.e., main ideas and supporting details, inferences [single-passage], author's craft, vocabulary [interpreting words and phrases in context, and decoding and recognizing words*], synthesis [paired argumentative passages])

Content Revision and Editing for Conventions Strand

- Essay Revision and Editing (i.e., development, organization, effective language use, and standard English conventions)
- Sentence Revision, Editing, and Completion (i.e., conventions of grammar; conventions of usage; conventions of punctuation; conventions of spelling and capitalization*; purpose and organization*; and sentence combining*)

Note: The Text Analysis and Synthesis strand may not initially contain many Literary Text Analysis sets with difficulties corresponding to diagnostic level 2 (i.e., Beginning Basic)[20]. As part of the postlaunch effort, College Board will evaluate the feasibility of adding literary sets appropriate for these levels.

A single testing experience consists of 48 questions, covering both reading and writing, across two strands. Questions are presented in a seamless experience, with no break or division between reading-focused and writing-focused diagnostic questions. Reading-focused questions appear first, beginning with three Literary Text Analysis sets; writing-focused questions follow, beginning with three Essay Revision and Editing sets.

The Text Analysis and Synthesis strand of the ELAR Diagnostic Test consists of 24 questions:

- 3 4-question Literary Text Analysis sets
- 12 discrete Informational Text Analysis and Synthesis questions
    - 2 Synthesis (paired argumentative passages) questions
    - 10 questions algorithmically chosen among all remaining Informational Text Analysis varieties

The Content Revision and Editing for Conventions strand contains 24 questions:

- 3 4-question Essay Revision and Editing sets
- 12 discrete questions algorithmically chosen among all Sentence Revision, Editing, and Completion varieties

The following tables provide synopses of key aspects of the ELAR CRC and Diagnostic Tests. Table 3.1 presents test content distributions on the ELAR CRC and Diagnostic Tests, and Table 3.2 presents a fuller articulation of ELAR CRC and Diagnostic question content.

[20] TSIA2 diagnostic levels are closely aligned to NRS Educational Functioning Levels.

Table 3.1:
TSIA2 ELAR CRC and Diagnostic Test Content Specifications

	Content Areas	Number of questions	Percentage of test
ELAR CRC			
Reading-focused	Set-based questions • 1 Literary Text Analysis set	4	13.3
	Discrete questions • 2 Synthesis (paired argumentative passages) • 9 Informational Text Analysis	11	36.7
Writing-focused	Set-based questions • 1 Essay Revision and Editing set	4	13.3
	Discrete questions • 11 Sentence Revision, Editing, and Completion	11	36.7
ELAR CRC Total		**30**	**100.0**
ELAR Diagnostic			
Text Analysis and Synthesis strand	Set-based questions • 3 Literary Text Analysis sets	12 (4 per set)	25.0
	Discrete questions • 2 Synthesis (paired argumentative passages) • 10 Informational Text Analysis	12	25.0
Strand Total		**24**	**50.0**
Content Revision and Editing for Conventions strand	Set-based questions • 3 Essay Revision and Editing sets	12 (4 per set)	25.0
	Discrete questions • Sentence Revision, Editing, and Completion	12	25.0
Strand Total		**24**	**50.0**
ELAR Diagnostic Total		**48**	**100.0**

Table 3.2:
ELAR CRC and Diagnostic Test Question Content

Reading-focused (CRC); Text Analysis and Synthesis (Diagnostic)		
Literary Text Analysis	The student will identify and analyze ideas in and elements of literary text.	
	Explicit information	The student will identify ideas explicitly stated and clearly indicated in literary text.
	Inferences	The student will draw reasonable inferences from literary text.
	Author's craft	The student will analyze an author's word choice rhetorically; analyze text structure, purpose, and audience; and analyze point of view and perspective in literary text.
	Vocabulary	The student will determine the meaning of words and phrases in context in literary text.
Informational Text Analysis and Synthesis	The student will identify and analyze information and ideas in and elements of informational text.	
	Main ideas and supporting details	The student will identify main ideas of and comprehend explicitly stated and clearly indicated information and ideas in informational text.
	Inferences (single-passage)	The student will draw reasonable inferences from informational text.
	Author's craft	The student will analyze word choice rhetorically; analyze text structure, purpose, and audience; and analyze point of view and perspective in informational text.
	Vocabulary (interpreting words and phrases in context; decoding and recognizing words*)	The student will determine the meaning of words and phrases in context in informational text and (Diagnostic only) apply decoding and word recognition skills.
	Synthesis (paired argumentative passages)	The student will draw reasonable connections between two related argumentative texts, including determining rhetorical relationships, analyzing commonalities, and analyzing claims and counterclaims.

Table continues

Writing-focused (CRC); Content Revision and Editing for Conventions (Diagnostic)		

Table 3.2 continued

Essay Revision and Editing	The student will revise and edit prose text as needed to improve development, organization, and language use as well as to ensure conformity to the conventions of standard written English grammar, usage, and punctuation.	
	Development	The student will revise as necessary to improve the development of text.
	Organization	The student will revise as necessary to improve the organization of text.
	Effective language use	The student will revise as necessary to improve the precision, concision, and context appropriateness of expression.
	Standard English conventions	The student will edit text as necessary to ensure conformity to the conventions of standard written English grammar, usage, and punctuation.
Sentence Revision, Editing, and Completion	The student will edit and complete sentences as necessary to ensure conformity to the conventions of standard written English grammar, usage, punctuation, and (Diagnostic only) spelling and capitalization as well as make effective decisions regarding purpose and organization (Diagnostic only) and sentence combining (Diagnostic only).	
	Conventions of grammar	The student will edit and complete sentences as necessary to ensure conformity to the conventions of standard written English grammar.
	Conventions of usage	The student will edit and complete sentences as necessary to ensure conformity to the conventions of standard written English usage.
	Conventions of punctuation	The student will edit and complete sentences as necessary to ensure conformity to the conventions of standard written English punctuation.
	Conventions of spelling and capitalization*	The student will edit and complete sentences as necessary to ensure conformity to the conventions of standard written English spelling and capitalization.
	Purpose and organization*	The student will make effective decisions regarding the appropriateness of written material for a given purpose and/or audience and the organization of written material.
	Sentence combining*	The student will combine two sentences into an effective single sentence.

* Content included on Diagnostic Test only

Key Features of the Multiple-Choice TSIA2 ELAR CRC and Diagnostic Tests

Key features of the TSIA2 ELAR Tests include

- Words in context
- Specified range of text length
- Specified range of text complexity
- Command of evidence
- Diversity

Words in context. Some questions on the ELAR tests measure test takers' understanding of the meaning and use of words and phrases in the context of prose passages. These words and phrases are neither highly obscure nor specific to any one domain. Instead, they are words and phrases whose specific meaning and rhetorical purpose are derived in large part through the context in which they are used.

Text length. Some questions use stimuli made up of single sentences, while others use passages of various lengths. Passage length is determined by a standard word count formula in which a *word* is defined as six characters. Most passages used for reading-focused questions range in length from 40 to 400 standard words, with one passage on the CRC Test in the range of 350 to 500 standard words. Writing passages range in length from single sentences to prose passages of up to 350 standard words.

Text complexity. Passages used in both the reading- and writing-focused elements of the ELAR Tests exhibit a defined range of text complexity from early high school level to postsecondary entry. To ensure that texts are appropriately challenging, test development staff use qualitative measures of text complexity as well as feedback from secondary and postsecondary subject matter experts and test data on student performance. The computer-adaptive test design, to some extent, influences the distribution of text complexity encountered by any given test taker. The qualitative text complexity rubric can be found in Appendix A: Text Complexity (Qualitative)—Reading and Writing.

Command of evidence. Questions associated with writing-focused prose passages in the ELAR Tests measure test takers' capacity to revise a text to improve its development of information and ideas. To answer such questions, test takers must have a solid grasp of the content of the passage (although, importantly, prior knowledge of the topic isn't expected or assessed).

Diversity. College Board is committed to presenting students with a test-taking experience that is reflective of the diversity of the United States and the world. To that end, passage and question pools include substantial content that visibly reflects U.S.-based racial and ethnic diversity (including African American/Black, Native American, Asian American, and Latinx individuals, cultures, and experiences); international/global (non-U.S.) perspectives, cultures, and settings; and balanced representation of genders.

Essay Test

In conjunction with the multiple-choice ELAR Tests, TSIA2 offers an Essay Test, which remains unchanged from TSIA1. Used in conjunction with the multiple-choice ELAR Tests to classify test takers into college ready or not college ready categories with respect to reading and writing, the Essay Test is intended to collect evidence in support of a broad claim about student performance:

Students can demonstrate college readiness proficiency in writing.

In its standard form, the Essay Test is delivered via and scored by computer. A range of accommodated versions are available for test takers with documented disabilities that may prevent them from taking the computer-delivered assessment.

Task format. The Essay Test consists of a single constructed-response prompt.

Stimulus content. The Essay prompt includes a brief text (*passage*) for test takers to read and consider as they develop their written response to the question (*assignment*). This passage is not a reading passage per se: test takers are not assessed on their comprehension of this passage, nor do they need to discuss it in their response. Rather, it serves primarily as "food for thought" and to contextualize the assignment.

The assignment that makes up part of the prompt is the question that test takers' response is intended to address (e.g., "Are we free to make our own decisions, or are we limited in the choices we can make?"). In response, test takers are asked to write an essay of 300 to 600 words. The essays, electronically scored, are evaluated on the test takers' developed ability to produce writing that emphasizes precise use of language, logical presentation of ideas, development of a point of view, and clarity of expression—traits highly valued in college courses.

Test taker responses, scored on a holistic rubric, are evaluated on six dimensions. Table 3.3 provides a synopsis of the dimensions assessed.

Table 3.3:
Essay Test Dimensional Score Descriptions

Dimension	Description
Purpose and Focus	The extent to which the writer presents information in a unified and coherent manner, clearly addressing the issue
Organization and Structure	The extent to which the writer orders and connects ideas
Development and Support	The extent to which the writer develops and supports ideas
Sentence Variety and Style	The extent to which the writer crafts sentences and paragraphs demonstrating control of vocabulary, voice, and structure
Mechanical Conventions	The extent to which the writer expresses ideas using Standard English conventions
Critical Thinking	The extent to which the writer communicates a point of view and demonstrates reasoned relationships among ideas

For an in-depth look at the Essay scoring rubrics, see Appendix C: TSIA2 Essay Scoring Rubrics.

Key Features of the Essay Test

Key features of the Essay Test include

- Skills that matter most
- Accessible contexts
- Accessible language

Skills that matter most. Students taking the TSIA2 Essay Test are scored on their command of writing skills required in most college courses. These skills include planning and presenting information and ideas in a coherent piece of writing, using precise language, presenting ideas logically, developing a point of view, and expressing ideas clearly.

Accessible contexts. As a writing test for students representing a wide range of cultural and linguistic backgrounds, experiences, and lengths of exposure to U.S. culture and English-speaking environments, the TSIA2 Essay Test uses contexts carefully developed to be accessible and free of elements that would impede access to the content. Essay prompts are made up of short passages, do not test specialized, technical, or literary topics, and do not require knowledge of U.S. culture or norms.

Accessible language. Every care is taken to ensure that prompts assess writing achievement and that the reading required to understand and fulfill the task is as clear and unambiguous as possible. No specialized, technical, or literary language is used to elicit test takers' writing samples.

Mathematics CRC Test

The Mathematics CRC Test is designed primarily to classify test takers into college ready or not college ready categories with respect to mathematics. The test consists of 20 questions and is intended to collect evidence in support of a broad claim about student performance:

> *Students can demonstrate college readiness proficiency in mathematics.*

In its standard form, the CRC Test is delivered adaptively via computer. A range of accommodated versions are available for test takers with documented disabilities that may prevent them from taking the computer-delivered assessments; tests in these formats are fixed-form linear (i.e., not adaptive).

Question format. All CRC questions are multiple-choice and discrete.

Question content. Test questions cover four main categories, where each covers a range of topics:

- Quantitative Reasoning
  - Compare magnitudes of rational and irrational numbers
  - Solve problems with ratios, proportions, and percentages
  - Solve proportional relationship problems in context (e.g., linear relationships in financial literacy and numeracy)
  - Identify, manipulate, and interpret linear equations, inequalities, and expressions
- Algebraic Reasoning
  - Solve linear equations, inequalities, and systems of linear equations
  - Evaluate linear functions
  - Solve quadratic and exponential relationship problems in context (e.g., exponential decay/growth, compound interest, and depreciation)
  - Identify and manipulate quadratic, polynomial, exponential, rational, and radical equations and expressions
  - Solve equations and evaluate functions (e.g., quadratic, polynomial, exponential, rational, and radical)
- Geometric and Spatial Reasoning
  - Convert units within systems of measurement
  - Find perimeter, area, surface area, and volume using a variety of methods, including estimation
  - Use transformations to investigate congruence, similarity, and symmetry
  - Apply right triangle relationships and basic trigonometry
  - Make connections between geometry and algebraic equations

- Probabilistic and Statistical Reasoning
  - Compute and interpret probability
  - Compute and describe measures of center and spread of data
  - Classify data and construct appropriate representations of data
  - Analyze, interpret, and draw conclusions from data

A single testing experience consists of 20 questions:

Quantitative Reasoning	6 questions
Algebraic Reasoning	7 questions
Geometric and Spatial Reasoning	3 questions
Probabilistic and Statistical Reasoning	4 questions

Test Summary. Tables 3.4 and 3.5, below, present synopses of key aspects of the Mathematics CRC and Diagnostic Tests.

Mathematics Diagnostic Test

The Mathematics Diagnostic Test is designed primarily to identify test takers' academic strengths and weaknesses with respect to mathematics. In its standard form, the Diagnostic Test is delivered adaptively via computer. A range of accommodated versions are available for test takers with documented disabilities that may prevent them from taking the computer-delivered assessments; tests in these formats are fixed-form linear (i.e., not adaptive).

Question format. All diagnostic questions are multiple-choice and discrete.

Question content. Just like the CRC Test, the Diagnostic Test consists of questions that cover the same four main categories, where each covers a somewhat wider range of topics. An asterisk (*) below denotes content on the Diagnostic Test not found on the CRC Test.

- Quantitative Reasoning
  - Perform basic mathematics operations with whole numbers and integers, decimals, and fractions*
  - Round numbers to a given decimal place*
  - Compare numbers in a variety of forms, including decimals, fractions, and percentages*
  - Compare magnitudes of rational and irrational numbers
  - Solve problems with ratios, proportions, and percentages
  - Solve proportional relationship problems in context (e.g., linear relationships in financial literacy and numeracy)
  - Identify, manipulate, and interpret linear equations, inequalities, and expressions

- Algebraic Reasoning
  - Solve linear equations, inequalities, and systems of linear equations
  - Evaluate linear functions
  - Solve quadratic and exponential relationship problems in context (e.g., exponential decay/growth, compound interest, and depreciation)
  - Identify and manipulate quadratic, polynomial, exponential, rational, and radical equations and expressions
  - Solve equations and evaluate functions (e.g., quadratic, polynomial, exponential, rational, and radical)
- Geometric and Spatial Reasoning
  - Identify common units of measurement*
  - Identify and define types of angles*
  - Convert units within systems of measurement
  - Find perimeter, area, surface area, and volume using a variety of methods, including estimation
  - Use transformations to investigate congruence, similarity, and symmetry
  - Apply right triangle relationships and basic trigonometry
  - Make connections between geometry and algebraic equations
- Probabilistic and Statistical Reasoning
  - Sort and count data*
  - Construct simple graphs and tables*
  - Compute and interpret probability
  - Compute and describe measures of center and spread of data
  - Classify data and construct appropriate representations of data
  - Analyze, interpret, and draw conclusions from data

A single testing experience consists of 48 questions, 12 questions per strand, across the four strands. The following tables provide synopses of key aspects of the Mathematics CRC and Diagnostic Tests. Table 3.4 presents test content distributions on the Mathematics CRC and Diagnostic Tests, and Table 3.5 presents a fuller articulation of Mathematics CRC and Diagnostic question content.

Table 3.4:
TSIA2 Mathematics CRC and Diagnostic Test Content Specifications

Content Areas	Number of questions	Percentage of test
Mathematics CRC		
Quantitative Reasoning	6	30
Algebraic Reasoning	7	35
Geometric and Spatial Reasoning	3	15
Probabilistic and Statistical Reasoning	4	20
Mathematics CRC Total	**20**	**100**
Mathematics Diagnostic		
Quantitative Reasoning	12	25
Algebraic Reasoning	12	25
Geometric and Spatial Reasoning	12	25
Probabilistic and Statistical Reasoning	12	25
Mathematics Diagnostic Total	**48**	**100**

Table 3.5:
Mathematics CRC and Diagnostic Test Question Content

Content Category	Content Subcategory	Description
Quantitative Reasoning	Perform basic mathematics operations with whole numbers and integers, decimals, and fractions*	The student will use mathematical symbols to represent words that represent those symbols, and add, subtract, multiply, and divide whole numbers, decimals, and fractions.
	Round numbers to a given decimal place*	The student will round to a specified place value, including 10, 100, and 1,000.
	Compare numbers in a variety of forms, including decimals, fractions, and percentages*	The student will compare and order whole numbers, decimals, and fractions (including on a number line).
	Compare magnitudes of rational and irrational numbers	The student will apply mathematical operations to rational and irrational numbers.
	Solve problems with ratios, proportions, and percentages	The student will apply ratios, proportions, and percentages to solve problems.

Table continues

Content Category	Content Subcategory	Description
		Table 3.5 continued
Quantitative Reasoning	Solve proportional relationship problems in context (e.g., linear relationships in financial literacy and numeracy)	The student will formulate a solution to a real-world situation based on the solution to a mathematical problem.
	Identify, manipulate, and interpret linear equations, inequalities, and expressions	The student will recognize and use algebraic properties, concepts, procedures, and algorithms to combine, transform, evaluate, and interpret expressions and equations.
Algebraic Reasoning	Solve linear equations, inequalities, and systems of linear equations	The student will recognize and use algebraic properties, concepts, procedures, and algorithms to solve equations, inequalities, and systems of linear equations, as well as make connections among graphical, tabular, and algebraic representations.
	Evaluate linear functions	The student will evaluate a linear function for a particular value.
	Solve quadratic and exponential relationship problems in context (e.g., exponential decay/growth, compound interest, and depreciation)	The student will formulate a solution to a real-world situation based on the solution to a mathematical problem.
	Identify and manipulate quadratic, polynomial, exponential, rational, and radical equations and expressions	The student will recognize and use algebraic properties, concepts, procedures, and algorithms to combine, transform, and evaluate expressions and equations.
Geometric and Spatial Reasoning	Solve equations and evaluate functions (e.g., quadratic, polynomial, exponential, rational, and radical)	The student will recognize and use algebraic properties, concepts, procedures, and algorithms to solve equations and evaluate functions, as well as make connections among graphical, tabular, and algebraic representations.
	Identify common units of measurement*	The student will identify length, area, volume, time, and temperature as standard measurements.
	Identify and define types of angles*	The student will identify and define angles, including supplementary, complementary, and vertical angles.
	Convert units within systems of measurement	The student will use proportional reasoning to convert units of measurement.
	Find perimeter, area, surface area, and volume using a variety of methods, including estimation	The student will recognize, identify, and validate properties of two- and three-dimensional figures, as well as calculate perimeter, area, surface area, and volume of figures.
		Table continues

Content Category	Content Subcategory	Description
		Table 3.5 continued
Geometric and Spatial Reasoning	Use transformations to investigate congruence, similarity, and symmetry	The student will identify and apply transformations to figures.
	Apply right triangle relationships and basic trigonometry	The student will apply right angle relationships and use basic trigonometric ratios to solve problems.
	Make connections between geometry and algebraic equations	The student will make connections between geometry and algebra.
Probabilistic and Statistical Reasoning	Sort and count data*	The student will sort and count data.
	Construct simple graphs and tables*	The student will construct simple graphs and tables to represent data.
	Compute and interpret probability	The student will compute and interpret the probability of an event and its complement.
	Compute and describe measures of center and spread of data	The student will compute and describe summary statistics of data.
	Classify data and construct appropriate representations of data	The student will identify appropriate representations of data based on its type.
	Analyze, interpret, and draw conclusions from data	The student will determine types of data, analyze given information, and determine a solution.

* Content included in Diagnostic Test only

Key Features of TSIA2 Mathematics CRC and Diagnostic Tests

Key features of the TSIA2 Mathematics Tests include:

- Mathematics that matters most
- Calculator and no-calculator questions
- Problems grounded in real-world contexts
- Specified range of text length
- Connection to mathematics pathways

Mathematics that matters most. The TSIA2 Mathematics Tests focus on knowledge and skills that are essential for college and career readiness, according to Texas's own curriculum and standards. These include an emphasis on applied reasoning questions over reasoning questions disconnected from the mathematics curriculum as well as a strong emphasis on both fluency with mathematical procedures and conceptual understanding.

Calculator use. The Mathematics Tests include questions without a calculator option as well as questions with one or more calculator options. No-calculator questions assess fluency in rational number arithmetic and include conceptual questions for which a calculator is not needed, while questions with calculator options give insight into students' capacity for strategic use of the tool to address problems efficiently.

Problems grounded in real-world contexts. The Mathematics Tests include a proportion of contextualized questions allowing test takers to engage with issues related to work performed in college and career and to mitigate the disconnect between mathematics concepts and real-life applications.

Specified range of text length. College Board test developers take pains to make sure that contextualized questions measure the relevant mathematical construct only and not reading skills or knowledge. Questions in the mathematics pool fall into three word-count levels: low (i.e., fewer than 40 words; these include questions that are not contextualized), medium (i.e., 40–60 words), and high (i.e., more than 60 words). The assembly of the pool ensures that the majority of questions have low or medium word counts.

Connection to mathematics pathways. Compared to the TSIA1 Mathematics Tests, the TSIA2 Mathematics Tests have a stronger and clearer connection to mathematics pathways. For example, quantitative reasoning now constitutes its own broad content category (one of four) in both the CRC and Diagnostic Tests. There is also more emphasis on reasoning skills throughout. These are deliberate efforts to help prepare Texas students for a range of college majors and career paths as well as for productive engagement in a society and economy that are increasingly reliant on data and quantitative reasoning.

3.3 Development of TSIA2 Assessments

Question Specifications

Except for the Essay Test, which elicits student writing using prompts, all questions in TSIA2 are multiple-choice, with each question having one and only one correct or best answer. Many multiple-choice questions are discrete in format, while some in the ELAR tests are set-based. All essay prompts present a similar level of challenge, providing a consistent and reliable measure of test taker achievement.

In keeping with AERA/APA/NCME Standard 4.7 (AERA, APA, & NCME, 2014), the following section describes how College Board creates and reviews the multiple-choice questions and prompts in our assessment programs, including TSIA2.

Crafting of Questions and Tasks

According to AERA/APA/NCME Standard 4.0, "Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of test scores for their intended uses" (AERA, APA, & NCME, 2014, p. 85). To this end, College Board measurement and assessment staff, in consultation with both the THECB and Texas educators, created the TSIA2 test design and question/task

specifications to represent the depth and breadth of the defined domains. The specifications define the question/task types and formats required to measure most directly and authentically the domains of skills and knowledge relevant to TSIA2's primary purposes and the overall claims.

Using final test specifications, College Board test developers began the process of assembling the TSIA2 question pools. First, developers identified a large group of questions that matched the defined domains, skills, and knowledge assessed in TSIA2 and have performed well in TSIA1 and other existing ACCUPLACER test pools. This large group of questions was then presented to external content experts with assessment development and teaching experience for initial review. Prior to reviewing the questions, these reviewers were trained on the TSIA2 test specifications, test purpose, College Board's fairness standards, and review process. They then independently reviewed and coded questions for inclusion or exclusion. College Board test developers led several meetings during the review period to ensure that the external content experts remained aligned with one another in their reviews. The external content experts' final selections and recommendations were considered by College Board's test developers, who then made one of three decisions for each question reviewed: 1) use the question in TSIA2 in its existing form, 2) edit and re-pretest the question before adding it to the question pool, or 3) reject the question for use in TSIA2.

From test design to pool assembly, the entire process of creating TSIA2 was informed by feedback from the THECB and Texas content experts.

Question Content and Fairness Reviews

TSIA2 measures skills and knowledge needed in postsecondary education, work, and life, specifically as these are described in the curriculum and standards[21] that inform the tests. As mentioned previously, College Board content and measurement staff worked closely with the THECB and academic committees convened by the THECB. These educators and content experts from across Texas included high school teachers and postsecondary instructors of entry-level courses, some of whom have experience in adult basic education. This collaboration has helped ensure that the questions and tasks in TSIA2 are aligned with Texas's best classroom practices.

Question Content

In order to consistently develop assessments with engaging, rich stimulus materials and contexts that lend themselves to high-quality questions and tasks, College Board has developed and continues to maintain a range of internal test support materials intended to help ensure that all questions and tasks meet or exceed industry standards and best practices, as defined principally by the AREA/APA/NCME *Standards for Educational and Psychological Testing*. These materials include question/task writer content and fairness guidelines as well as question/task prototypes and templates. To develop and/or review all questions and tasks, College Board contracts with faculty and educational professionals at

[21] Texas College and Career Readiness Standards (2018); Texas Essential Knowledge and Skills (TEKS), English III (2017), Algebra II (2012); AEL Content Standards 2.0; and NRS EFL.

both the high school and postsecondary levels and with other independent content and instructional experts. In this way, those most familiar with the student populations and knowledgeable in best instructional practices make a significant contribution to assessment content. This contribution helps ensure that the test materials included in the assessments are engaging, instructionally appropriate, and fair to all students.

Multiple-choice questions

ELAR. In the reading-focused elements of the ELAR CRC and Diagnostic Tests, students engage with texts worth reading and worthy of careful consideration. Some passages are selected from previously published authentic writing that exemplifies the genres represented on the test; others are commissioned passages of high quality. The essential first step of reading-focused question development is a close and careful reading of the focal text. Reading-focused test questions resemble questions that might emerge naturally in a thoughtful classroom conversation and return students to the text to examine closely the information and ideas within it. The best test questions develop out of a sensitive engagement with the passage rather than an effort to try to cover in a mechanical way every possible testing point in the domain. Such questions also favor a more organic development process that respects the unique natures of rich texts in a variety of content areas.

The writing-focused elements consist of passages that are engaging and challenging, paired with questions that focus clearly on a core set of writing and language requirements. These commissioned passages are designed to provide meaningful contexts for the skills and knowledge being addressed and to exemplify the qualities of effective arguments, informative/explanatory texts, and nonfiction narratives. Some questions assess writing and language skills and knowledge in extended prose contexts, while others ask test takers to read and identify errors in single-sentence stimuli.

Mathematics. The Mathematics CRC and Diagnostic Tests ask test takers to demonstrate their command of the mathematics most provably useful in a range of college courses and career environments. They provide the opportunity for richer applications of the most essential mathematics to address situations and problems grounded in the real world.

Test questions are thoughtfully designed with the help of educators with a deep knowledge of the target mathematical content and practices. The questions on each Mathematics Test emphasize the use of mathematics in unlocking insights and solving problems. The test design allows the core of mathematics to be examined with the range of rigor required (as defined through evidence) for college and career readiness, assessing at once students' procedural skill, application ability, and conceptual understanding. Rather than covering a broad number of topics that most students will never see again, the Mathematics Tests encourage students to study fewer topics that represent a deep core that they can draw on again and again in their schooling, college, and career, as represented in the standards that inform TSIA2. At the same time, the assessments include pure mathematics problems that focus on the type of reasoning essential for success in solving diverse problems and engaging in demanding disciplines.

Essay Test

Essay Test prompts are written by test developers who are subject-matter specialists, many of whom have either high school or college teaching experience. The prompts are then reviewed by writing faculty members representing high schools and two- and four-year colleges from around the United States. Each prompt is written to be easily accessible to a wide test-taking population, including students from a range of age groups and for whom English is a second language.

The Essay Test is intended to give test takers the opportunity to use a broad range of experiences, learning, and ideas to support their point of view on the issue presented. Prompts do not draw on specialized knowledge in any particular area or on any specific course material that a student may have studied; they are likewise free of figurative, technical, or specific literary language or references. Contexts are reflective of a range of student interests, including the arts, sports, technology, science, and history.

Each prompt presents a short passage that stimulates critical reflection and allows test takers to draw on their knowledge and interests to respond. The passage is no more than 80 words and is followed by a writing assignment that focuses the test taker on the issues addressed in the passage. Passages are typically based on previously published texts and selected based on their utility for the task and appropriateness and suitability for a wide audience. As much as possible, such passages are kept intact as they originally appeared in publication, although they may be minimally adapted to provide greater accessibility for students (e.g., to ensure fairness or eliminate unduly obscure or difficult vocabulary or construction). Other passages are written for the test; these are developed to meet the requirements just described, including passage length and appropriateness for the target audience. Each passage is followed by a prompt-specific assignment that succinctly states the issues presented in the passage and identifies possible points to consider as students plan and write their essay.

All prompts are written to meet a set of criteria: they must present an issue that will engage test takers from a broad range of backgrounds and allow them to draw on their knowledge and interests to respond; they must stimulate critical reflection on the issue by suggesting a range of possible viewpoints on it (i.e., passages may present opposing points of view on an issue, but each side should be complex enough to allow the student to develop a variety of positions within each point of view); they must avoid moralizing statements that might encourage socially desirable or hypocritical responses not reflective of test takers' true opinions; they must not be flat statements of fact; and they must not ask students to address a particular audience (e.g., their high school principal) or write a certain kind of response (e.g., a letter to the mayor), as such requirements might make test takers think there is one socially or politically appropriate way to address the prompt, thus potentially impeding the writing activity these prompts are developed to elicit.

Content and Fairness Review

In keeping with AERA/APA/NCME Standard 3.2 (AERA, APA, & NCME, 2014), test developers work to ensure that test materials, including multiple-choice questions and Essay Test prompts, are fair to all

students. This work involves (1) ensuring that scores are not influenced by construct-irrelevant factors, such as test takers' gender or race/ethnicity; (2) confirming that materials are appropriate for the broad and diverse test-taking population; (3) avoiding or eliminating sensitive subject matter that might adversely affect performance of test takers as a whole or members of certain population groups; and (4) identifying and addressing instances in which test materials are likely to advantage or disadvantage members of particular population groups due to construct-irrelevant factors (i.e., avoiding or eliminating bias). To supplement the steps taken by College Board staff, all test materials undergo external review by independent panels of educators, assembled to reflect the diversity of test takers, for any issues that would affect the fairness of test materials. These panels are made up of high school and college faculty reflecting a variety of academic disciplines, geographic regions, genders, and races/ethnicities. As a result of internal and external fairness review, test materials may be accepted, revised to address problematic content, or discarded. For more information on fairness reviews prior to field testing, see Chapter 2 of this manual, specifically Section 2.3: Fairness of TSIA2 Assessments.

Question Pretesting, Analysis, and Calibration

Question Pretesting

Multiple-Choice Questions. Every operational question in TSIA2 has previously been pretested; that is, the question has been embedded in an operational test and administered (not for a score) to students in the target population to make sure that the question is not ambiguous or confusing and to determine the difficulty level and the degree to which it differentiates among higher- and lower-achieving test takers. The pretest responses are also analyzed to determine whether test takers of different racial/ethnic groups or genders, having similar achievement levels, respond to the question differently.

For each pretested question, the data from at least 1,000 test takers are used to evaluate question performance. This information provides an accurate estimate of how the question will function when administered operationally.

Essay Prompts. In keeping with AERA/APA/NCME Standard 4.8 (AERA, APA, & NCME, 2014), it is important to ensure that the prompts "function similarly for different groups" (p. 88). After prompt reviews and a resolution process to address any concerns raised during the reviews, new Essay prompts are field tested with a representative sample of test takers in a special administration in classrooms around the country. For each group of prompts field tested, a diverse sample of schools is invited to participate by having students respond to a particular prompt. The students who participate in field testing vary by race/ethnicity, gender, and socioeconomic status.

The responses gathered from field tests are read by a group of experienced writing instructors to determine whether a particular prompt is readily understood by test takers and elicits responses that reflect differing degrees of writing skill. Members of this group individually read and score a substantial number of the responses. As a group, they discuss each prompt and decide whether it is usable, needs revision, or should be discarded. From the student responses collected during the field testing, exemplars are chosen for each point on the holistic scoring scale. These serve as anchor papers for

training essay readers when the essay prompt is administered operationally. The scoring process is described in more detail in Chapter 5: Interpretation and Application of Results.

Analysis of Pretest Information

In keeping with Standard 4.10, data collected from pretests are analyzed to provide important information about the appropriateness of questions to be included in the question pools of the TSIA2 suite (AERA, APA, & NCME, 2014). An initial item analysis is performed on the data to provide test developers with statistical information to review questions for any possible issue with regard to keys and distractors, as well as an alert for possible issues that will affect question or item calibrations. The main statistics computed are indices of difficulty, discrimination, and differential item functioning (DIF).

Question Difficulty and Discrimination. Initial analysis of difficulty and discrimination is based on Classical Test theory (CTT). In CTT, question difficulty is the percentage of test takers who answer the question correctly. It is typically referred to as the "p-value" of the question. A high p-value indicates an easy question; that is, a question that most of the test takers answered correctly. A low p-value indicates a hard question, one that most of the test takers answered incorrectly. Question discrimination is the correlation between the score on that question and the total score on the test. Questions that correlate well with total test score tend to correlate well with one another and produce a test that is more reliable. A high correlation indicates that the question performs as expected, in that the proportion of higher-scoring test takers answering the question correctly is greater than the proportion of lower-scoring test takers answering correctly. A low correlation indicates a question not performing as intended and requires a review by test development experts and may need to be removed from the test.

When a question is dichotomously scored, point-biserial correlation is equivalent to well-known Pearson correlation coefficient to indicate question-total correlation coefficient to indicate question-total correlation. The computation of the point-biserial index is shown in the equation below. There, $\overline{x}_{i1}$ is the mean scale score for test takers who answer question i correctly, $\overline{x}_{i0}$ is the mean scale score for test takers who answered question i incorrectly, p_i is the proportion of test takers that answered question i correctly, and S_x is the standard deviation of scale scores.

$$r_{pbs} = \frac{(\overline{x}_{i1} - \overline{x}_{i0})\sqrt{p_i(1-p_i)}}{S_x}$$

In addition to the difficulty and discrimination indices above, other statistics computed at this stage are:

- The number and percentage of test takers who selected each distractor
- The point-biserial correlation for each distractor
- The average scale score for test takers who selected each distractor

CTT question statistics are used to flag questions for a closer examination. The following are the criteria used to flag questions for further content review:

- Question Difficulty < 0.15; Question Difficulty > 0.90
- Question Discrimination < 0.10

Questions are also flagged based on the performance of the following distractors:

- Distractor Discrimination: > 0.05
- Distractor Attracting < 1% of all test takers
- Distractor Average Scale Score Higher than that for the Keyed Response Option

Questions with incorrect key, unclear distractors, extreme p-values, or low question-total correlations are dropped from calibration.

Differential Item Functioning (DIF)

Establishing the fairness of tests is an important part of supporting and justifying the use of test scores for their intended purposes. Of particular concern in establishing test fairness is ensuring that questions are equally informative for different subgroups of test takers. For test scores to be valid, it is important to be certain that there is nothing influencing responses to questions other than the knowledge, skills, and abilities (KSAs) that the questions on the test intend to measure (Zumbo, 1999). Anything unrelated to the intended KSAs that differentially influences the responses of subgroups of test takers is a threat to the validity of score interpretations. In short, when subpopulations of test takers are matched on their abilities, there should be no difference in their achieving a particular score on test questions or items. To identify any potentially unfair test questions, a statistical technique known as Differential Item Functioning (DIF) analysis can be employed.

DIF is a statistical observation that involves matching test takers from different groups on the characteristic measured and comparing performance across groups on each question. Test takers of equal ability who belong to different groups should respond similarly to a given test question. If they do not, the question is said to function differently across groups and is classified as a DIF question (see Clauser & Mazor, 1998, or Holland & Wainer, 1993 for more complete descriptions of DIF theory and methodology). Differential performance alone does not mean a question is biased. Bias is present when a question has been statistically flagged for DIF and the reason for the DIF is traced to a factor irrelevant to the construct the question is intended to measure. Therefore, for a question to be considered biased, a characteristic of the question that is unfair to one or more groups must be identified. TSIA2 test questions flagged for DIF are sent to test developers for review.

For analysis of DIF for gender, the performance of male test takers is compared to the performance of female test takers, with the males serving as the reference group. For analysis of DIF for ethnic/racial groups, the performance of White test takers as the reference group is compared to other ethnic/racial

subgroups. Ethnicity is defined as Hispanic or non-Hispanic, and race is defined as American Indian or Alaska Native, Asian, Black or African American, Multiple Races, and White. All non-Hispanic respondents are identified as one of the previously listed racial categories. The minimum sample size requirements are 50 for the focal group and 100 for the reference group when calculating the statistics.

There are many methods for detecting DIF (Clauser & Mazor, 1998; Camilli & Shepard, 1993; Holland & Wainer, 1993). Most of them are not applicable for computer-adaptive tests (CAT), since CAT doesn't have the same set of questions, or even the same length of the test, for all test takers. Logistic regression (LR) is one of the methods that is appropriate for CAT DIF detection for the following advantages (Sireci, 2001; Swaminathan & Rogers, 1990; Zumbo, 1999). First, LR does not require test takers to take the same set of questions. Second, LR can detect both uniform and non-uniform DIF. Uniform DIF occurs when the probability of getting a question correct is higher for one group across the ability level. Nonuniform DIF occurs when the probability of getting a question correct is higher at one range of ability level but lower at the other range for one group. Third, simulation studies have shown that LR has acceptable power and type I error rate when employing the effect size measure (Jodoin & Gierl, 2001). DIF analyses for tests in the TSIA2 suite employ Zumbo's 1999 method.

In this method, three models are constructed for each question. Accordingly, the R^2 , which is based on the likelihood ratio χ^2 for testing the null hypothesis that all coefficients are 0 (Cox & Snell, 1989), can be obtained for each. The effect sizes for uniform and non-uniform DIF are $R_2^2 - R_1^2$ and $R_3^2 - R_2^2$, respectively. The models and effect sizes are in in Table 3.6. For the TSIA2 suite, an effect size is considered:

- negligible if it is less than 0.034
- moderate if it is greater than or equal to 0.034 and less than 0.07
- sizeable if it is greater than or equal to 0.07.

Questions with sizeable effect sizes, those questions favoring one group over the other for test takers of the same ability, are not included in the calibration. Questions with severe DIF are automatically removed from the question pool, while those with moderate or negligible DIF are retained for use unless internal and/or external content review identifies one or more construct-irrelevant factors likely contributing to the DIF results. If the reviewers determine that the DIF is due to a factor irrelevant to the construct the test is supposed to measure, the question is considered to be biased; such questions are either revised and repretested (and again analyzed for DIF) or removed. Note that for a question to be biased, at least one characteristic of the question that is unfair to one or more population groups must be identified.

Table 3.6:
Models and Effect Sizes

Index	Model	R^2	Effect Size
1	$y = \beta_0 + \beta_1 theta$	R_1^2	
2	$y = \beta_0 + \beta_1 theta + \beta_2 group$	R_2^2	$R_2^2 - R_1^2$
3	$y = \beta_0 + \beta_1 theta + \beta_2 group + \beta_3 theta * group$	R_3^2	$R_3^2 - R_2^2$

Question Pool Calibration

Questions that are not dropped from the question pool based on the initial item analysis and subsequent review are included in the question calibration. Question or item calibration is the term commonly used to describe Item Response Theory (IRT) item parameter estimation. IRT is fully described in the CAT Algorithm section of this chapter and it is suggested that, as the details of IRT and parameter estimation are beyond the scope of this manual, interested readers refer to Hambleton and Swaminathan (1985), Hambleton, Swaminathan, and Rogers (1991), Lord (1980), Lord & Novick (1968), and Baker and Kim (2004). Excellent discussions of IRT within the context of computerized adaptive testing can be found in Wainer (2000).

IRT calibrations for the TSIA2 ELAR and mathematics questions were performed using FlexMIRT® Version 3.51 (Cai, 2017). After an initial calibration, item model-fit was inspected. In some cases, questions were eliminated prior to final calibration. The item parameters and the item response functions were examined for abnormalities. Questions not rejected after final calibration became part of the respective question pools.

3.4 Computer-Adaptive Test Algorithm

The TSIA2 suite is administered as a computer-adaptive test (CAT), which allows for instantaneous score reporting. The technology in CAT affords the capability to provide accurate and efficient measurement of a test taker's knowledge and skills. As soon as a test taker finishes a test, his or her Individual Score Report is available and is immediately exportable into existing campus information systems.

The previous sections in this chapter discussed the creation and testing of questions to be administered in TSIA2 CRC and Diagnostic Tests. As the last few sections discussed the establishment of a question pool, this section will explain the workings of the CAT algorithm for the TSIA2 suite. In keeping with AERA/APA/NCME Standard 5.16, "when test scores are based on model-based psychometric procedures, such as those used in computerized adaptive or multistage testing, documentation should be provided to indicate that the scores have comparable meaning over alternate sets of items" (AERA, APA, & NCME, 2014, page 106).

As in any CAT, the adaptive algorithm used for the TSIA2 tests is designed to arrive at reliable scores as efficiently as possible. The rationale for this approach is that it is unnecessary for test takers of high ability to take the easiest questions or for test takers of low ability to take the hardest questions, as doing so does not contribute much to the quality of the ability estimates. Questions with difficulty levels that are far away from a test taker's ability level do not contribute enough information about estimation of that test taker's ability to be of practical use. Choosing questions that contribute information is more efficient and provides more accurate scores.

In a typical CAT, a test taker is initially presented with a question of a designated difficulty level. In some testing programs, the first question is of medium difficulty, while in TSIA2 the first question is of a slightly lower difficulty to allow a positive introduction to the testing experience. If the test taker's response to the first question is correct, a more difficult question is presented next. If the response is incorrect, the test taker is presented with a less difficult question. An ability estimate based on the test taker's previous responses is computed after each response and successive questions are presented to meet content specifications and to provide as much information as possible about the test taker's ability. The test is terminated after the specific number of questions are administered.

A typical CAT system consists of several components:

- An item response theory (IRT) model
- A calibrated test question pool
- An initial trait level to begin the test
- A procedure for selecting test questions
- A method for estimating ability
- A criterion to terminate the test

The rest of this section is organized according to these components.

Item Response Theory (IRT) Model

Constructs such as ability in a subject matter are not directly observable. The responses that test takers provide to questions related to the construct to be assessed provide information about the unobserved construct. The test development process, from defining the construct of interest to generation, review, and pretesting of the questions, ensures that the questions presented operationally to a test taker are appropriate representations.

Tests are sets of questions that are an operationalization of the construct of interest. IRT models are intended to relate the estimated response to the underlying construct. IRT models describe a probabilistic relationship between a test taker's response to a test question and some latent trait such as mathematics, reading, or writing ability. Test takers with higher ability have a higher probability of answering a question correctly than test takers with lower ability. Figure 3.1 depicts this relationship; as

test takers increase in ability, as indicated on the X-axis, the probability of answering the question correctly increases, as indicated on the Y-axis. The figure is an Item Characteristic Curve (ICC) for a specific test question based on the three-parameter logistic model that is used for the TSIA2 CRC and Diagnostic Tests.

With the exception of the Essay Test, all questions in the TSIA2 suite are multiple-choice. For such tests, a dichotomous model is most appropriate. The most general of the common dichotomous models is the three-parameter logistic model. The three parameters are discrimination (known as a-parameter), difficulty (known as b-parameter) and pseudo-guessing (known as c-parameter). Referring to Figure 3.1, the a-parameter is proportional to the slope of the ICC at the difficulty level of the question (b, discussed below). The steeper the slope the more discriminating the question is performing around that difficulty level. That is, with a higher slope it does not require a large change in ability to increase the probability of answering correctly. The probabilities of answering a question correctly increase more slowly for a less-discriminating question. Typically, it is desirable to have questions with a-values of 1 or higher, but content constraints and the difficulty of creating questions with high discriminations at differing ability levels generally means that questions with a-values lower than 1 are often used.

The pseudo-guessing parameter, c, represents the probability of answering the question correctly for test takers with extremely low or no knowledge of the construct. For multiple-choice tests, it is possible to answer a question correctly by guessing.

The difficulty parameter, b, occurs at the ability level where the test takers with that level have a certain probability of answering correctly. If there were no guessing, the b-parameter would be at the ability level where there was a 50% chance of answering the question correctly. When there is possibility of guessing, the b-parameter would be at the ability where the probability of answering the question correctly is equal to one-half of the sum of 1 and the guessing parameter c; $(1 + c_i)/2$.

Theta (θ) is the ability level on the underlying and unobservable trait being measured. The range of θ is theoretically from negative infinity (absolutely no knowledge) to positive infinity (perfect knowledge). Though the scale for ability and difficulty parameters is arbitrary, most IRT software scales test taker parameters so that 0 is the average, and that the standard deviation of the abilities is generally set to 1. These default values were used for TSIA2 ELAR and mathematics calibration. This means that a question with a b-parameter of 0 is usually considered to be of average difficulty. For TSIA2 calibration, the range of θ is between -6 and 6.

The three-parameter model is represented by the following equation:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability of a correct response to question i, given an ability level of θ. The question or item parameters are a_i, b_i, and c_i and refer to characteristics of the questions themselves.

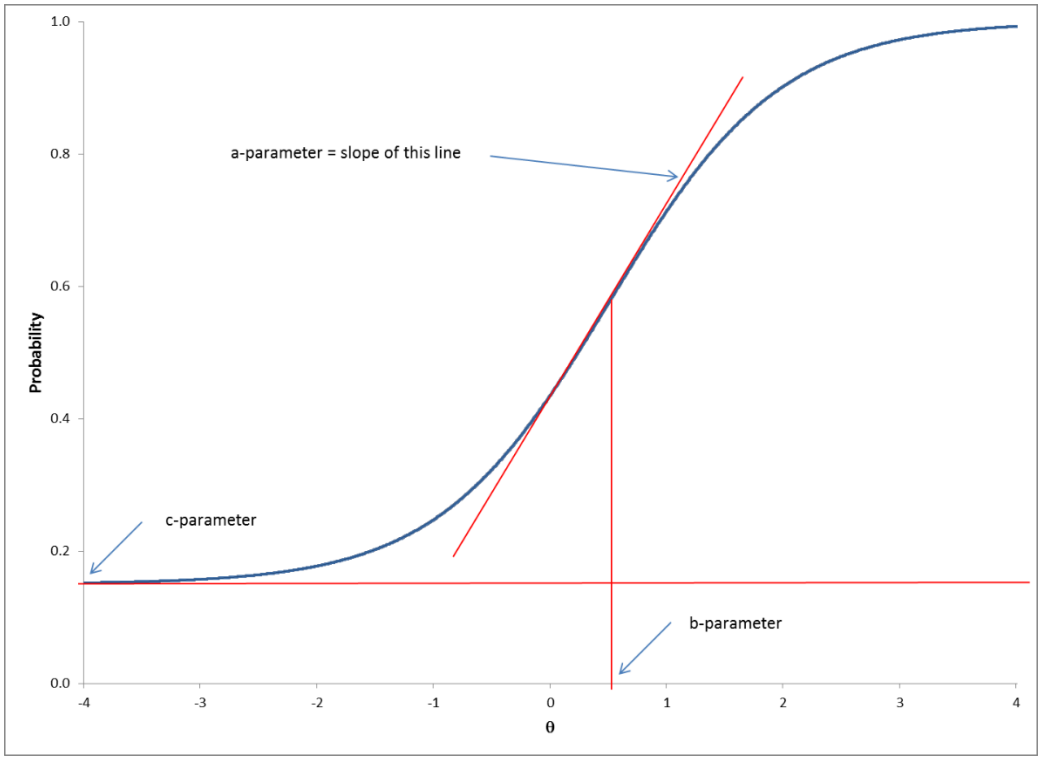

Figure 3.1. Graphical Representation of Item Characteristic Curve

Other popular IRT models are special cases of the more general three-parameter logistic model. The two-parameter model is the case where the c-parameter is set equal to zero. The one-parameter model is obtained when the c-parameter is zero and the a-parameter is set equal to 1 for all questions.

The calibrations performed with IRT have the result that all test takers' θ values and all question or item parameters are on scale. That is, by virtue of all the questions being on the same scale, each test form (i.e., selection of questions administered to student) is effectively pre-equated. Thus, post-equating (which is what most people call equating) is unnecessary.

Calibrated Question Pool

With requirements and constraints that need to be satisfied for a successful administration of a CAT, a rich question pool is essential. For each non-essay test in TSIA2, a pool of test questions written to the various test content areas is developed and, following extensive review including pretesting, calibrated to the selected IRT models. The number of questions on each test question pool is presented in Tables 3.7 through 3.8 below. The distribution of questions across content areas covered in each test is also presented. Other statistical characteristics of the test question pools are presented in Appendix D: Statistical Characteristics of TSIA2 Test Question Pools.

Table 3.7:
TSIA2 ELAR Test Question Pool Content Distribution

Content/Strand	CRC Test		Diagnostic Test	
	N	%	N	%
Reading-Focused	187	69.52	603	54.03
Literary Text Analysis	32	11.90	92	8.24
Informational Text Analysis and Synthesis	155	57.62	511	45.79
Writing-Focused	82	30.48	513	45.97
Essay Revision and Editing	32	11.90	104	9.32
Sentence Revision, Editing, and Completion	50	18.58	409	36.65
Total	269	100.00	1,116	100.00

Table 3.8:
TSIA2 Mathematics Test Question Pool Content Distribution

Content/Strand	CRC Test		Diagnostic Test	
	N	%	N	%
Quantitative Reasoning	62	19.14	241	33.66
Algebraic Reasoning	124	38.27	278	38.83
Geometry and Spatial Reasoning	72	22.22	79	11.03
Probability and Statistical Reasoning	66	20.37	118	16.48
Total	324	100.00	716	100.00

Initial Trait Level

For each CRC Test, the first question is chosen based on a relatively low initial ability estimate of $\theta = -1.0$ to allow most test takers a successful experience in the beginning of the test. From the second question on, the next question administered to a test taker is automatically chosen based on the skill level indicated by answers to all previous questions and the content specifications that are still to be met.

Question Selection Procedure

Each TSIA2 test is tailored to the test taker using a question selection algorithm that takes into account content balance, measurement precision, and question exposure control. Content balancing is an important consideration; it ensures that different tests across test takers cover the same proportion of content categories so that test takers are measured on the same composition of traits. The adaptive nature of the tests involves identifying and administering the questions in the pool that provide the

most information at the current estimate of ability for each test taker as they progress through the test. Question exposure control is another important practical consideration. Because CATs are continuously administered from the same question pool over a period of time, some "popular" questions may become known and no longer provide valid measurement. To ensure appropriate content coverage, efficient and accurate scores, and to prevent overexposure of questions, the TSIA2 CAT system incorporates statistical algorithms to control content balancing, information, and question exposure rate into the question selection process.

The exposure control algorithm used for TSIA2 tests is based on the Conditional Randomesque method (Kingsbury & Zara, 1989).This method allows a preset maximum exposure rate for a specific ability range so that exposure will be constrained at various ability levels. For CRC and Diagnostic Tests, the preset maximum exposure rate is 0.25.

The Conditional Randomesque strategy randomly selects the next question to be administered from the group of the most informative questions, given the current θ estimate. The selection of the question is always made at random among the most informative questions. The Conditional Randomesque repeatedly selects the same number of the most informative questions (e.g., 2, 3, 4…10) from which one is randomly selected for administration throughout testing and does not switch to maximum information selection at any time. Kingsbury and Zara suggest that continuing the randomization technique throughout testing will decrease the overlap in questions seen by test takers of similar abilities. The number of most informative questions from which one will be selected for administration to the test taker is 4 for all TSIA2 tests. These group sizes were found to work well for preset maximum exposure rate of 0.25.

To ensure content balance on the TSIA2 tests, a number of constraints are built into each test with respect to content category and inter-item dependency. Inter-item dependencies deal with relations of exclusion and inclusion between questions in the pool (Veldkamp & van der Linden, 2000). As part of inter-item dependency, constraints can be specified for "enemy" questions such that the presentation of one question will preclude the appearance of another question in the enemy list on the same test, or for "set" questions such that all questions in the same set will be administered together on the test.

To balance content, the TSIA2 CAT algorithm uses the adjusted Weighted Penalty Function method (Segall and Davey, 1995; Fan, 2007) which takes into account the content-related constraints, question information, and sufficiency of questions related to each constraint. This method assigns each eligible question in the pool a penalty value at each question selection level, with questions having smaller penalty values deemed more desirable for selection. After the penalty value for each question is computed, a list of questions with the lowest penalty values is formed and provided to the question exposure control method for further question selection. This ensures that the next question to be administered is selected according to content constraints, maximum test information, and the desired question exposure rate. The content categories for each test and the number of questions selected from each category are presented in Tables 3.1 and 3.4 earlier in this chapter.

Ability Estimation Procedure

Although fewer questions are presented with a CAT for each test than would be given in a linear test, greater accuracy is maintained on trait estimates by providing challenging tests that correspond to each test taker's ability level, and by using an ability estimation algorithm that ensures accurate and efficient ability estimates. For ability estimation, the Maximum Likelihood Estimation procedure that employs the Newton-Raphson method and Brute Force method (Birnbaum, 1968; Hambleton & Swaminathan, 1985) is implemented in the TSIA2 CAT algorithm. The Newton-Raphson method is the most commonly used method in IRT pattern scoring; it has significant advantage in efficiency but has two misuses that can occur for some test takers. Some sequences of question responses from a test taker may be inconsistent with expectations (e.g., answering some harder questions correctly and easier questions incorrectly), resulting in an inability of the method to converge on an estimate of ability. For other patterns, the method may find a local maximum estimate when, in fact, another real maximum exists elsewhere on the ability scale. The Brute Force method is an algorithm often used in Computer Science for searching sets of data for an answer. It requires more extensive computation but guarantees a true maximum likelihood solution. For the TSIA2 tests, the Newton-Raphson method is used as the main ability estimation method, with the Brute Force method as a supplement when the convergence and local maximum issues occur. Estimated abilities are in the -5.0 to 5.0 range of scores. The reported scores for each are based on a linear transformation from the ability estimates. How the linear transformation is derived is discussed in Chapter 6: Psychometrics, specifically Section 6.1: Scaling.

Termination Criterion

Multiple-choice TSIA2 CRC and Diagnostic Tests are *fixed-length* CAT tests, and their questions are dichotomously scored. The number of questions on each test is presented in the Test Specifications section of this chapter (Section 3.2). In a *variable-length* CAT, which is used by some testing programs, the test terminates when the precision of the ability estimate for a test taker reaches an established threshold. Because TSIA2 test have a predetermined number of questions to be administered for each test, a variable-length CAT termination rule was not used.

Other Features of the CAT Algorithm

The TSIA2 test administration platform allows pretest questions, which do not count toward test takers' scores, to be embedded in operational CAT tests for the purpose of gathering response data. In addition to the operational CAT algorithm, there is a CAT simulation system that has the exact same functionalities as the operational system. The simulation system can simulate CATs and analyze the characteristics of the simulated tests to provide information about the performance of the CAT algorithm. For the current CAT algorithm, extensive simulation studies have been conducted; these studies indicate that the current system produces tests that meet content and question exposure requirements sufficiently well and provide ability estimates that are psychometrically efficient and accurate, resulting in valid and reliable test scores.

3.5 Accommodations

TSIA2 Accommodations

In keeping with AERA/APA/NCME Standards (AERA, APA, & NCME, 2014), College Board believes that "all test takers should have the full opportunity to demonstrate their standing on the construct being measured" (p. 52). Consistent with the Americans with Disabilities Act and to ensure fairness across assessments, students who present documentation that their disabilities affect their ability to participate in TSIA2 are eligible for accommodations. Approval from College Board is not required to administer accommodations on any TSIA2 test; institutions work with their Services for Students with Disabilities (SSD) coordinators to determine eligibility based on test takers' documented disabilities.

COMPANION Forms

College Board is committed to making assessments available in accessible formats. All tests in TSIA2 have two corresponding, comparable COMPANION forms. COMPANION forms present TSIA2 content in alternate formats and are designed for test takers who are not able to take computer-adaptive tests or for institutions that may be unable to administer them. COMPANION forms are nonadaptive, linear tests that have been designed to proportionally align in content to the corresponding computer-adaptive tests. They use the same score scale as the computer-adaptive tests and, in terms of number of questions, are 1.5 times the length of their corresponding computer-adaptive tests. Like other tests in the TSIA2 suite, they are also untimed.

Several COMPANION formats are available: regular and large print "print-on-demand" test forms that test administrators may download from the platform; reader scripts, audio CDs; and braille. Table 3.9 shows the lengths of COMPANION forms relative to the computer-adaptive tests.

Table 3.9:
Number of Questions on Computer-Adaptive and COMPANION Tests

Test	Number of questions	
	Computer-adaptive	COMPANION
ELAR CRC	30	44
ELAR Diagnostic	48	72
Mathematics CRC	20	30
Mathematics Diagnostic	48	72

COMPANION forms yield inferences that are comparable to those from the computer-adaptive version of the assessment.

Development of COMPANION Forms

TSIA2 COMPANION forms have been developed to provide alternate, nonadaptive, linear formats to test takers or institutions that may be unable to access the corresponding computer-based tests. AERA/APA/NCME Standard 8.3 (AERA, APA, & NCME, 2014) states that "when the test taker is offered a choice of test format, information about the characteristics of each format should be provided" (p. 134).

COMPANION forms are designed to have content specifications that align with the corresponding computer-adaptive tests and to provide good measurement across the score scale. Two parallel COMPANION forms are available to test takers for every test. The two versions of each COMPANION form have similar reliabilities, which is to say that both forms are built to have comparable Test Information Functions (TIF). Figures E1-E10 in Appendix E: Test Information Function and Test Characteristic curves of COMPANION forms show the TIF and Test Characteristic Curves (TCC) of each COMPANION form.

COMPANION forms are created using questions selected from active question pools. Using an Automated Test Assembly (ATA) program described in Chuah, Hare, Bay, & Proctor (2020), the College Board Psychometrics group makes the initial selection of questions to be included in each test form. The draft forms, which meet content and statistical specifications, are then sent to College Board's Assessment Design and Development (AD&D) group for review. The AD&D review may result in a decision to replace a question or questions based on professional judgment. AD&D sends feedback to the Psychometrics team. Any changes needed, including question replacements, are made by Psychometrics. After verifying that content and statistical specifications are still met following any changes, the form is then sent back to AD&D for additional review. This process is repeated until the COMPANION forms are deemed final. During this process, AD&D also determines the order of the questions on the final test forms.

Development of Conversion Tables for COMPANION Forms

The raw scores that are computed for the COMPANION tests are the sum of the correct answers, the "number right." These raw scores are then converted to the reported score scale which is 910 to 990 for the ELAR and Mathematics CRC tests, and reporting categories for the diagnostic tests.

The conversion tables for the COMPANION forms were developed in order for the "number correct" scores from each COMPANION form to be placed on the appropriate score scale and/or performance categories. The analyses involved simulation studies using Item Response Theory (IRT) parameters for the questions of each COMPANION form. The IRT parameters are described further in the section of this chapter on CAT Algorithm (Section 3.4). A sample of 50,000 simulees with a uniform distribution of ability (theta) ranging from very low ability (theta equal to -5.00) to very high ability (theta equal to 5.00) is used for the simulation in order to ensure uniform coverage across the ability range.

First, these sampled thetas are used as true theta values as input to the simulations, and the simulation steps are as follows:

1. For each COMPANION form, test taker response data, which are strings of incorrect (0) and correct (1) responses, are simulated for each simulee via the three-parameter item response theory (IRT) model using the theta values and question or item parameters for the questions on the COMPANION test.
2. The total "number correct" score is then computed for each simulee on each form of each test by summing the zeroes and ones simulated in step 1. The "number right" ranges from zero to the total number of questions on the test. Note that scores as low as zero are generally not observed in simulations or in operational administrations.
3. The theta for each response string is then estimated using expected a posteriori (Bock & Mislevy, 1982) in IRT.
4. At each number correct score from zero to number of questions on the test, the estimated theta values for all simulees achieving that score are summarized and the average estimated theta is obtained.
5. The average estimated theta is then mapped to the scale score using the theta-to-scale score conversion table established for the computer-adaptive ELAR and mathematics CRC tests. For diagnostic tests, the average estimated theta is mapped to the proficiency level or NRS EFL using the conversion tables developed for their computer-adaptive test counterparts.

In summary, when a student takes the COMPANION forms of CRC tests, their number correct scores are mapped to a scale score on the 910 to 990 range. If a CRC score is below the college readiness benchmark, the student takes the appropriate diagnostic test. The total number correct on the diagnostic test is mapped to the corresponding NRS EFL. Furthermore, for each content strand the number correct is mapped to a corresponding proficiency level of Basic, Proficient, or Advanced.

Chapter 4 — Administration of TSIA2

Introduction

College Board works to ensure that all test scores are valid for their intended uses and that all test takers have a fair testing experience. This chapter documents the appropriate use of tests in the Texas Success Initiative Assessment 2.0 (TSIA2) suite, how they should be administered, and the steps all administrators of these tests must take to protect test materials and prohibit the inappropriate sharing of test information at any time.

TSIA2 is administered in the ACCUPLACER test administration platform. Access to the test administration platform is granted to nationally or internationally accredited, degree-granting institutions and systemwide educational governance.

Section 4.1 of this chapter discusses the appropriate use of TSIA2, designed to assess test takers' readiness for college-level coursework in the general areas of English language arts and reading (ELAR) and mathematics, as an assessment to help place students in higher education courses. It also highlights scenarios in which it is not appropriate to use TSIA2 as an assessment. Section 4.2 discusses the policies and procedures involved in TSIA2 test administration, including the computer-adaptive algorithm and the COMPANION forms. Section 4.3 discusses test security and the ways to prevent attempts to gain an unfair advantage and compromise test scores for their intended uses, including the environment in which these tests should be administered and the eligibility requirements and responsibilities of those administering the tests. We discuss these procedures as they apply to test materials and test takers, as well as the rationale behind these procedures.

4.1 Appropriate Use

TSIA2 was designed and developed specifically to assess the academic knowledge and skills of entering undergraduate students in ELAR and mathematics. The College Readiness Classification (CRC) tests were designed to be administered to all entering undergraduate students, and the scores from these tests are intended to be used for college placement purposes. For students scoring below the Texas Higher Education Coordinating Board (THECB)-designated benchmark for entry into college-level courses, the Diagnostic Tests are intended to be used for identifying specific areas of strength and weakness and to facilitate entry into the appropriate developmental education course or to support co-enrollment in a developmental education course and an entry-level, credit-bearing course within the same semester.

The administration of TSIA2 to high school students is appropriate to the extent that scores from the tests are used to determine college readiness and to connect students who are not college ready with appropriate interventions. These interventions are intended to equip students as quickly and efficiently as possible with the knowledge and skills needed to become college ready by no later than the end of high school.

4.2 Test Administration

Administration of Online Tests

All TSIA2 tests, except for the Essay Test, are computer adaptive. Computer-adaptive testing (CAT) is a mode of test administration that uses computer algorithms to select and deliver test questions to test takers. Test questions are selected from an operational pool that has been developed to provide optimal coverage for the various content areas specified for each test. The CAT algorithm is discussed in greater depth in Section 3.4 of this manual.

Each TSIA2 test question has been calibrated for difficulty and other characteristics. Unlike many traditional tests where all test takers take a single form of an assessment, the sequence of test questions and the questions themselves will vary from test taker to test taker. The next question administered to a test taker is automatically chosen to yield the most information about the test taker based on the skill level indicated by answers to all prior questions. The criteria for selecting the next question to be administered to a test taker are complex; however, the primary criterion is a desire to match the difficulty of the question to the test taker's current estimated ability.

The TSIA2 test delivery system adapts or "tailors" the test to each test taker by keeping track of a test taker's performance on each test question and using an item—or question—selection algorithm based on a weighted deviations model to determine the next question to be administered. During testing, the first question presented is of medium difficulty and is chosen randomly from several starter questions of the same level of difficulty. If a test taker answers the question incorrectly, the next question to be administered is chosen from a group of easier questions. If the test taker answers the question correctly, the next question presented will be somewhat more difficult.

The test delivery system continues this process throughout the test, choosing the next question that is expected to yield the most information about the test taker. To ensure that the test is balanced in content, and that the kinds of questions presented do not differ greatly from one test taker to another except in difficulty, several constraints are built into the program. These constraints guide the selection of questions to be administered so that a balance is achieved regardless of the skill level of the individual.

Administration of Accommodated Tests

COMPANION Tests provide accommodated formats for test takers who are unable to take computer-adaptive TSIA2 tests. COMPANION Tests are available for all TSIA2 tests. In addition to the regular paper-and-pencil format, COMPANION Tests are also available in braille, large print, and reader script formats. Audio CDs that test takers can use to hear TSIA2 test stimuli, questions, and answer choices are also available.

The COMPANION Tests typically have approximately 1.5 as many questions as the corresponding computer-adaptive tests. Questions for the COMPANION tests are selected using an Automated Test Assembly (ATA) program described in Chuah, Hare, Bay, & Proctor (2020). The program follows the same

content specification for the adaptive tests, resulting in test forms that are proportionally equivalent in content coverage. The program also uses statistical specifications to ensure that selected questions cover the difficulty range, from easy to difficult. Additional information on COMPANION test administration can be found in the TSIA2 Administrator's Manual.

4.3 Security

Institutions using the ACCUPLACER platform for administering TSIA2 are required to sign a License Agreement that requires all testing be done in a secure and proctored setting.

The License Agreement requires that:

- All testing be done in a secure and proctored setting
- Test takers be monitored at all times during a test session
- An authorized, certified test administrator from the institution or from a College Board approved remote proctoring vendor be present on-site or online during all administrations of online TSIA2 or COMPANION Tests
- Only approved users may log in to the platform to administer an assessment
- Test takers will not be permitted to log in to the platform on their own
- Under no circumstances should proctor login credentials be shared with test takers; and
- Login credentials may not be written on chalkboards or whiteboards, printed, emailed, or presented online in any form or place.

Test Center Guidelines

Before administering TSIA2 tests, administrators should evaluate their testing facilities and review testing procedures so as to ensure a comfortable, positive, and efficient testing environment and experience for test takers.

Below are guidelines for any TSIA2 testing environment:

1. The testing room must be appropriately heated or cooled, adequately ventilated, and free from distractions.
2. Lighting must enable all test takers to read the computer screen in comfort and should not produce shadows or glare on the computer screen or writing surfaces.
3. The testing room cannot contain maps, periodic tables, posters, charts, or any ancillary materials related to the subject matter of the tests.
4. The testing room must comfortably accommodate the number of testing stations placed in it.
5. Testing rooms must be quiet throughout the duration of each test administration. When testing is scheduled or is in progress, other activities that would disrupt the standardized testing environment cannot be conducted.

6. The building, testing rooms, and restrooms should be accessible to people with disabilities, including wheelchair accessibility.
7. Restrooms should be located near the testing room and should be easy to find. Post directional signs if necessary.
8. Unauthorized individuals (e.g., parents, chaperones, non-testing staff or students) are not permitted in the testing center during the test. Persons assisting for accommodation purposes (e.g., readers or scribes) are considered authorized.

Prohibited Items

The following items prohibited from the testing room:

1. Any nonmedical electronic devices, especially any device capable of recording audio, photographic, or video content or any device capable of viewing or playing back such content. This includes but is not limited to wireless communication devices such as cellular phones, tablets, pagers, smartphones, walkie-talkies, PDAs, digital cameras, digital watches, smartwatches, or wristwatch cameras; listening devices such as radios, media players (with or without headphones), or recorders; and flash/thumb drives or any other portable electronic storage or recording device.
2. Any unauthorized testing aids, including calculators (test takers with a prescribed accommodation and those taking an accommodated format exempted); test taker provided keyboard, computer, or laptop, unless there is a documented disabling condition that requires the use of such specific device; dictionaries (standard and/or bilingual), books, pamphlets, or other reference materials; and slide rules, protractors, compasses, or rulers.
3. Paper of any kind (scratch paper may be provided, and any scratch paper distributed for testing must be collected and destroyed by the test center administrator). For exams proctored via virtual remote proctoring, test takers can use their own scratch paper but must destroy it at the end of the exam in view of the proctor.
4. Food, beverages, or tobacco products.
5. Weapons, firearms, or other items prohibited by law or test center/campus safety and security policies.

Proctor eligibility and responsibilities

Proctors must meet the following eligibility criteria:

1. Proctors must review proctor training materials and pass the ACCUPLACER Certificate of Test Administration (ACTA).
2. Proctors must be responsible adults trained to administer standardized tests.
3. High school students are ineligible to proctor TSIA2.
4. Proctors must have their own username and password. Login credentials cannot be shared with anyone, including Institution Administrators (IAs) and Site Managers (SMs).
5. Proctors cannot administer a TSIA2 test to a member of their household, immediate and/or extended family members, or friends.

6. Proctors must not have a stake in the outcome of test takers' scores.
7. Proctors cannot be engaged with any commercial test preparation company. This includes employment, volunteering, consulting, or acting as independent contractors.
8. For any remote, off-campus location testing, proctors must be vetted and authorized by the affiliated institution to proctor assessments in such locations.

Proctors are eligible to receive proctor login credentials only after they have successfully passed the ACTA. These credentials are valid for one year and need to be renewed on an annual basis. IAs and SMs must select proctors who are trained in the administration of standardized tests, which includes how to safely secure all testing materials (online and COMPANION). In addition, IAs and SMs are expected to provide proctors with specific information about test administration procedures, as well as regular training. All parties involved in administering TSIA2 tests must adhere to the policies outlined in the ACCUPLACER License Agreement (found in the ACCUPLACER Program Manual at https://secure-media.collegeboard.org/digitalServices/pdf/accuplacer/accuplacer-program-manual.pdf).

Proctors must engage in active proctoring behavior. For example, they should circulate the testing room throughout the testing session to ensure that test takers are working on the correct test and not engaging in any prohibited behavior. During the administration of a test, proctors cannot engage in non-test administration activities such as reading, eating, drinking, conversing, or using cell phones or other electronic devices.

Proctor responsibilities vary and include the following:

1. Verifying the identity of every test taker before the administration of a test. In the event a test taker leaves the testing center for any reason during testing, identification must be re-verified upon their return to the testing center or upon receiving their Individual Student Report (ISR).
2. Collecting and/or storing test takers' unauthorized items (e.g., cellphones, smartwatches, and dictionaries) in a secure area that is not accessible to the test taker during the test. Test takers cannot place these within arm's reach (e.g., under their desks or chairs).
3. Supporting the IA and/or SM with securing all TSIA2 test materials.
4. Assisting test takers with testing equipment during testing and/or with logging in to the correct TSIA2 test.
5. Providing test takers with scratch paper and pencils and collecting and securely destroying all scratch paper once testing is completed. Test Takers cannot bring or use their own scratch paper (for exams proctored via virtual remote proctoring, test takers can use their own scratch paper but must destroy it at the end of the exam in view of the proctor.).
6. Printing and distributing ISRs to test takers after testing. Identification must be re-verified prior to providing an ISR to a test taker at the end of the test session.
7. Administering assessments to test takers with disabilities based on diagnosed accommodations.
8. Ensuring proper test security in advance of, during, and following testing sessions.

Chapter 5 — Interpretation and Application of Results

Introduction

To ensure that scores are usable for intended purposes, "assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation" (AERA, APA, & NCME, 2014, p. 114). Test administration procedures were discussed in Chapter 4: Administration of TSIA2. The first section of this chapter (Section 5.1) describes the scoring procedures for the Texas Success Initiative Assessment 2.0 (TSIA2) college readiness classification (CRC) and Diagnostic Tests and their COMPANION forms as well as the Essay Test. Because TSIA2 scores are used to make decisions on college readiness, procedures for setting college readiness benchmarks are discussed in Section 5.2. Test takers who score below the college readiness benchmarks take the Diagnostic Tests. To help students and institutions in interpreting performance in terms of strengths and weaknesses, proficiency statements for each Diagnostic Test strand are created. Proficiency statements, discussed in Section 5.3, provide a set of data-driven statements of what students know and can do at different ranges of performance on the diagnostic test strands. The chapter then covers reporting of results (Section 5.4).

5.1 Scoring Procedures

Computer-Adaptive Tests (CATs)

Generating test scores for a test taker involves several steps in a computer-adaptive test (CAT) and is covered in Chapter 3 of this manual. The test taker's ability estimate is computed, and an appropriate question chosen to be administered. This sequence is followed until the last question is administered. The process of choosing appropriate questions and calculating ability estimate is covered in Section 3.4: Computer-Adaptive Test Algorithm. The final ability estimate is based on the test taker's response string of 0s (incorrect responses) and 1s (correct responses). In other words, the ability estimate is computed based on which questions are selected for the test takers, which of those questions were answered correctly, and which questions were answered incorrectly. The ability estimate is then translated to the reported score scale ranging from 910 to 990, as discussed in Section 6.1: Scaling. Whether the scaled score is above or below the benchmark determines whether the test taker is college ready or routed to take Diagnostic Tests. The benchmarks are established as described in Section 5.2: Setting College Readiness Benchmarks.

COMPANION Forms

Each of the raw scores that are computed for the COMPANION forms is the sum of the correct answers or the "number right." Through simulation studies, an ability estimate corresponding to each raw score is computed. The transformation from ability estimate to scaled score used in CAT is also used for COMPANION forms. The simulation studies to determine the ability estimate for each raw score are described in Section 3.5 of this manual in the Development of Conversion Tables for COMPANION Forms.

Essay Scoring

The Essay Test measures test takers' ability to write effectively, which is critical to academic success. Each test taker's writing sample is scored on the basis of how effectively it communicates a whole message to readers for the purpose stated in the Essay prompt. The test taker's score is based on their ability to express, organize, and support opinions and ideas, not the position taken on the topic.

Each Essay Test gives test takers an opportunity to show how effectively they can develop and express their ideas in writing. They read a short passage and an assignment that are focused on an important issue and then write an essay in which they develop their own point of view on the issue. The passage is intended to stimulate thought about a topic or issue and test takers are asked to draw on a broad range of experiences, learning, and ideas to support their point of view on the issue in question.

Essay Tests are machine scored. Human participation only comes in at three steps:

- Scoring of essays used for initial calibration of the scoring engine
- Annual audit
- Checking of test taker essays that have anomalies and need to be referred to human readers.

College Board provides the Essay scoring guide, which is an eight-point holistic scale incorporating criteria that characterize writing. The writing sample is evaluated holistically. Essay scores range from 1 to 8. An essay that is too short to be evaluated, written on a topic other than the one presented, or written in a language other than English is given a zero. The Essay scoring rubrics can be found in Appendix C: TSIA2 Essay Scoring Rubrics.

In addition to holistic scores, additional feedback for each student essay is available in the form of dimension scores on:

- Purpose and focus
- Organization and structure
- Development and support
- Sentence variety and style
- Mechanical conventions
- Critical thinking.

Dimensional scoring rubrics are also available in Appendix C: TSIA2 Essay Scoring Rubrics. The score on each dimension has three levels. The dimensional scores are not meant to sum up to the holistic score for the Essay Test, they are only meant to provide additional feedback on the test taker's writing ability.

Technology Used to Score Essays

Essays are electronically scored by the Intelligent Essay Assessor (IEA) that is powered by the Knowledge Analysis Technologies (KAT) engine. Developed by the Knowledge Technologies group at Pearson, IEA is an automated assessment technology that evaluates the meaning of text, not just grammatical correctness or spelling.

IEA is based on Latent Semantic Analysis (LSA), a statistical language learning theory and computer model that measures the semantic similarity of words and documents with accuracy closely approximating that of human judges. LSA was originated at Bell Laboratories under Thomas Landauer, Ph.D., and was built into automated educational assessment products at the University of Colorado and Pearson.

IEA automatically evaluates the semantic substance of a test taker's writing by comparing a new essay to a set of essays that have each been graded by two expert human readers. IEA is able to do this comparison and produce accurate and reliable scoring because each essay prompt has been calibrated against 500 or more essays scored by human readers.

As a new essay is submitted, IEA looks for similarities to the scored essays and assigns a holistic score by placing it in a category with the essays to which it is most similar. Dimension scoring occurs in much the same way. For each dimension, the system assesses the submitted essay by comparing it to scored essays, and then categorizes the dimension in question. IEA includes built-in detectors for off-topic responses and other special situations that may need to be referred to human readers. The correlation and agreement rates of the scores produced by IEA have been shown to be as high as or higher than those between two independent readers.

Calibrating the Scoring Engine

Reader Training. Human readers score the essays that are used to calibrate IEA. As noted above, each essay prompt is calibrated against 500 or more test taker essays scored by human readers. Readers are people with at least a bachelor's degree who have been trained and undergone qualification to score the assessments. Readers selected to calibrate IEA all have previous successful scoring experience on at least one WritePlacer Annual Audit, preferably more than one. In addition to qualification statistics, agreement with prescored "validity" responses is used as a measure of audit scoring quality and is taken into consideration in the selection of high performing readers.

All readers attend an online training program that teaches the fundamentals of holistic and analytic scoring. The training is designed so that readers learn how to properly apply the rubrics. Readers must pass a baseline qualification test to score the Essay Test. They also receive training for every prompt that they score.

Baseline training encompasses one anchor set, two practice sets, and two qualifying sets. There are also three online training modules that every reader takes. Outlier readers also take a fourth module to train

how to score the types of outlier responses that cannot be scored by the KAT scoring engine as well as how to monitor workflow and manage their own scoring assignments to meet the 24-hour scoring deadline.

A key component of the training are anchor papers, which contain exemplar test taker essays that clarify the scoring guide and define the range that exists within each score point. The anchor papers demonstrate different approaches and different levels of achievement within each score point. Anchor sets are accompanied by annotations, which explain the score of each anchor paper using language from the rubric. Examples selected from the exemplar essays help explain the score. Ultimately, the annotation helps the reader understand not only that particular essay but also similar essays seen during training and scoring.

Readers are required to get at least 50% exact and 90% exact plus adjacent agreement on 1 of 2 qualifying sets on a baseline item. The qualifying standard is applied independently to each scoring trait. After qualifying, prompt-specific training is available for every standard Essay prompt. Prompt-specific training is 1 anchor set followed by a single practice set.

For Essay scoring, College Board uses a 60-day requalification rule. If a reader goes more than 60 days without scoring at least one essay, they must requalify.

Process for Obtaining Scored Anchor Essays for Training the Scoring Engine*.* Anchor papers from administration to administration are reused since this project uses the same prompts year to year. Anchor sets for all prompts were created using live responses from outlier scoring. Sets were built by content experts with over 10 years of experience working on writing assessments. Anchor papers are generated from live outlier essays.

Anchor sets cover the 8-point holistic score range as well as the dimension scores, but the primary training focus is on the holistic score. All anchor papers are annotated.

Human Reader Validation of Scores for a Sample of Essays*.* As a quality measure to ensure consistency of scores between human readers and the rubric, a variety of techniques to monitor scoring quality are implemented.

Backreading is a primary tool for proactively guarding against reader drift. The scoring system's integrated backreading tool allows their supervisory staff to review the scores assigned to individual test taker responses by any given reader.

Scoring directors and supervisors can perform a search for:

- Responses scored by a particular reader
- Responses receiving a particular score point
- Responses with scores that agree with, are adjacent to, or are nonadjacent to each other
- Combinations of these features.

Scoring directors use calibration sets to reinforce rangefinding standards, introduce scoring decisions, or correct scoring issues and trends. The primary goal of calibration is to continue training and to reinforce the scoring standards. Calibration sets may be "on the line" between score points or may contain unusual examples that are challenging to score and therefore useful for reinforcing the scoring rubric. After scoring an online calibration set, readers have an opportunity to ask questions of scoring supervisors and to seek clarification of the score point or annotation.

Scorer exception processing allows project managers to define intervals at which their scoring system checks for exact and adjacent agreement. If readers fall below preset standards, messages are automatically sent, interrupting their scoring process. Project leadership determines appropriate steps to remediate the reader. The reader may then work with a scoring supervisor, review anchor papers, or work through other activities to improve their scoring.

Through this process, the scoring system can automatically send an additional training/requalification set, and if performance is not improved, can lock readers out of the scoring system. This automated process complements Pearson's supervisory methods and prevents readers from continuing to score if standards are not maintained.

Validity essays are prescored essays strategically interspersed in the pool of live responses. These essays are not distinguishable from live essays and readers' scores are only accepted for monitoring purposes, not in replacement of the predetermined "true scores."

The validity mechanism provides an objective and systematic check of accuracy. It verifies that readers are applying the same standards throughout the project and, therefore, guards against reader drift and ultimately group drift. This procedure provides immediate feedback on individual readers and the group as a whole.

The validity pool includes responses encompassing the entire score range for each prompt. Readers score these responses without being aware that they are validity responses, which will provide informative statistical scoring information. Validity responses are sent to readers throughout the project.

Select validity responses are annotated by the scoring director and flagged for review. If a reader scores one of these responses incorrectly, the scoring session is interrupted while the response appears on the reader's screen with the true score, the score they're assigned, and an annotation. This immediate feedback greatly aids in preventing reader drift before it occurs. Once a reader has received feedback about a specific validity response, the response is flagged so the reader does not receive it again.

Interpreting Essay Results: Essay Dimensions

In addition to the reported holistic score, feedback is provided on six dimensions considered essential in a well-written essay (College Board, 2018).

Purpose and Focus – the extent to which the writer presents information in a unified and coherent manner, clearly addressing the issue. Specific elements to consider include:

- Unity
- Consistency
- Coherence
- Relevance
- Audience

Organization and Structure – the extent to which the writer orders and connects ideas. Specific elements to consider include:

- Introduction
- Thesis
- Body paragraphs
- Transitions
- Conclusions

Development and Support – the extent to which the writer develops and supports ideas. Specific elements to consider include:

- Point of view
- Coherent arguments
- Evidence
- Elaboration

Sentence Variety and Style – the extent to which the writer crafts sentences and paragraphs demonstrating control of vocabulary, voice, and structure. Specific elements to consider include:

- Sentence length
- Sentence structure
- Usage
- Tone
- Vocabulary
- Voice

Mechanical Conventions – the extent to which the writer expresses ideas using Standard English. Specific elements to consider include:

- Spelling
- Grammar
- Punctuation

Critical Thinking – the extent to which the writer communicates a point of view and demonstrates reasoned relationships among ideas. Specific elements to consider include:

- Clarity
- Depth
- Precision
- Logic
- Accuracy
- Fairness
- Breadth
- Relevance

If dimension statements have been selected to be reported on the Individual Score Report, one of the dimension statements shown below will be reported for each of the indicated dimensions. Each statement in Table 5.1 below describes the test taker's skill in the indicated writing dimension.

Table 5.1:
Essay Dimension Scores and Descriptions

Purpose and Focus
Your response shows a clear purpose and a consistent focus.
Your response does not fully communicate purpose, and focus may be inconsistent.
Your response lacks clear purpose and focus.
Organization and Structure
Your response demonstrates strong organization of ideas.
Your response demonstrates limited organization of ideas.
Your response demonstrates poor organization of ideas.
Development and Support
Your response is logically developed and well supported.
Your response has limited support for your ideas.
Your response needs additional ideas and support.
Sentence Variety and Style
Your response shows skillful control of sentence structure and style.
Your response shows inconsistent control of sentence variety, word choice, and flow of thought.
Your response shows limited ability to vary sentence length and apply appropriate vocabulary.
Mechanical Conventions
Your response shows strong control of mechanical conventions such as grammar, spelling, and punctuation.
Your response shows limited control of mechanical conventions such as grammar, spelling, and punctuation.
Your response shows poor control of mechanical conventions such as grammar, spelling, and punctuation.
Critical Thinking
Your response shows clear and reasoned analysis of the issue.
Your response shows limited clarity and complexity of thought.
Your response shows insufficient reasoning and lacks complexity of thought.

5.2 Setting College Readiness Benchmarks

As was established in the previous chapters, TSIA2 is primarily used to determine college placement for Texas students. Depending on how students score on the CRC Test(s), they may be assigned another test in the TSIA2 suite or placed into an appropriate course. Cut scores or benchmarks are essential in determining these placements. To this end, a standard setting session was convened by the Texas

Higher Education Coordinating Board (THECB) and College Board to determine the CRC benchmarks for the Mathematics and ELAR CATs. In keeping with AERA/APA/NCME Standard 5.21, which states that "the rationale and procedures used for establishing cut scores should be documented clearly" (AERA, APA, & NCME, 2014, page 107), the TSIA2 Standard Setting Report (Bay & Duffy, 2020) was submitted to the THECB for their decision making. What follows is a description of the standard setting process and results.

On July 21-22, 2020, content expert panelists for ELAR and Mathematics were gathered for the purpose of setting the college readiness benchmarks for the new TSIA2 tests. The level of preparation was commensurate with the rigor of implementation ensuring procedural validity and resulting in cut scores that are reliable, realistic, and useful.

A 21-member panel was assembled for each of ELAR and Mathematics. Panelists have varied levels of teaching experience, as shown in Table 5.2. Note that the categories of panelists' teaching experience are not mutually exclusive. Panelists were recruited by the THECB to ensure that all levels of experience were represented in the panel and that there is a diverse statewide representation.

Table 5.2:
Teaching Experience of Standard Setting Panelists

Teaching Experience	ELAR	Mathematics
College Level	17	15
Developmental Education	14	15
Adult Basic Education	3	4
High School	8	12

The Bookmark method (Mitzel, et al. 2001) was selected not only because is it deemed the most appropriate but also because it is considered the industry standard for educational assessment. This method is relatively easy to use and "perhaps the most popular method currently used to set performance standards on large-scale educational achievement test" (Cizek, 2012), and has withstood legal challenges (see, for example, Lewis, et al., 1999 & Mitzel, et al. 2001). Using this method, panelists reviewed a set of test questions that were ordered from the easiest to the most difficult. A bookmark is placed immediately preceding the question that they judge to be too difficult for a test taker who the panelist considers "barely" college ready. The median where the panelists placed their bookmarks is where each panel set their bookmark.

Because the standard setting had to be held during a pandemic, a previously planned in-person implementation was adjusted to an online implementation. Utmost care was observed so that the adjustments:

- Minimize the impact of change in mode from in-person to online
- Elicit the desired behavior of all those involved
- Maintain or enhance test security

For each of the ELAR and Mathematics Tests, two standard setting processes were implemented. Each process was implemented in two rounds. During the standard setting implementations, test taker performance was presented in the ability (i.e., θ) scale. Transforming the ability estimates from the θ scale to the reporting scale of 910 to 990 is discussed in detail in Chapter 6 of this manual.

The first implementation was to set the college readiness benchmark. On the reporting scale of 910 to 990, the final cut scores are 945 for ELAR and 950 for mathematics. This means that a test taker who scores 945 or higher in the multiple-choice test (and a score of 5 on the Essay) will be classified as college ready in ELAR. Similarly, a score of 950 or higher in mathematics will classify a test taker as college ready. Based on simulations, it is expected that 29.98% of test takers in ELAR and 21.92% of test takers in mathematics will earn the college ready classification. A test taker not deemed college ready after taking the initial tests will be routed to take the Diagnostic Test where they will have a second chance to be classified as college ready.

The second standard setting implementation was to set the cut scores on the Diagnostic Tests. The five diagnostic levels are consistent with the NRS levels 2, 3, 4, 5, and 6. Thus, for each Diagnostic Test four cut scores were to be set. Prior to standard setting, it was decided that test takers who perform at level 5 in ELAR will be classified as college ready. Similarly, test takers who perform at level 6 in mathematics will be classified as college ready. Thus, on the second standard setting implementation ELAR diagnostic cut scores for levels 3, 4, and 6 were set for ELAR and cut scores for levels 3, 4, and 5 were set for mathematics. The CRC cut score on the θ scale was held constant for level 5 in ELAR. Similarly, the CRC cut score on the θ scale was held constant for level 6 in mathematics. The final cut scores for the Diagnostic Tests are presented in Table 5.3.

Table 5.3:
Final Diagnostic Cut Scores

Diagnostic Level Cut Score	ELAR	Mathematics
2/3	-1.0285	-1.5363
3/4	-0.2740	-1.0581
4/5	0.7797	-0.1589
5/6	1.8336	0.5113

Standard setting results may generally be viewed in three parts:

1. Performance Level (or Borderline) Descriptors
2. Cut Scores
3. Evaluation Summary

Only the cut scores are presented here. Please refer to the TSIA2 Standard Setting Report (Bay & Duffy, 2020) for the rest of the results and other details of the implementation.

5.3 Proficiency Statements for Diagnostic Tests

When students take the Diagnostic Tests, they receive a diagnostic profile. The diagnostic profile consists of a classification into one of five diagnostic levels closely aligned to the National Reporting System Educational Functioning Levels, a proficiency level (Basic, Proficient, or Advanced) in each of the diagnostic strands, along with proficiency statements describing typical performance relative to the three tiers of achievement on each of the test strands. There are two diagnostic strands in ELAR and four in mathematics:

- ELAR
  - Text Analysis and Synthesis
  - Content Revision and Editing for Conventions
- Mathematics
  - Quantitative Reasoning (QR)
  - Algebraic Reasoning (AR)
  - Geometric and Spatial Reasoning (GSR)
  - Probabilistic and Statistical Reasoning (PSR)

For each strand, a statement of what test takers know and can do at each tier of achievement—Basic, Proficient, and Advanced—was created.

Proficiency statements are designed to help students gain a better understanding of how scores relate to specific academic skills. They offer descriptions of performance and insight into skills measured at

performance level. Proficiency statements provide a set of data-driven information intended to help students interpret their performance in ELAR and mathematics. They describe what a test taker scoring at the Basic, Proficient, or Advanced level is likely to know and be able to do in relation to the academic skills measured on the tests. Proficiency statements help students, teachers, administrators, and others understand what their Diagnostic Test score means and, for Basic and Proficient levels, how performance could be improved.

Determining Score Ranges for Proficiency Levels for each Diagnostic Strand

Determining the score ranges for the three levels of proficiency in each diagnostic strand would require setting cut scores that delineate adjacent levels. The COVID-19 pandemic limited the opportunities to implement a standard setting for the diagnostic strands. It is fortunate that TSIA2 has a very successful precursor program–TSIA1. An equipercentile linking (Kolen and Brennan, 2004) between corresponding diagnostic strands was performed to determine interim cut scores. Final cut scores may be determined through standards verification.

Equipercentile Linking

Using real data from TSIA1 collected in the 13 months prior to the pandemic, equipercentile linking was used to project the current cut scores on to the scale of corresponding tests on TSIA2 simulated data. For mathematics, the TSIA2 diagnostic strands (e.g., QR, AR, GSR, PSR) are generally aligned to the TSIA1 mathematics diagnostic strands Elementary Algebra (EA), Intermediate Algebra (IA), Geometry and Measurement (GM), and Data Analysis, Statistics, and Probability (DSP), respectively. Using the sample of test takers who took the TSIA1 diagnostic test, the cut scores were projected on to the TSIA2 mathematics ability scale using the sample from the simulated data with true ability that's comparable to the ability of the students who took the TSIA1 Mathematics DE Diagnostic Test.

The two strands in the TSIA2 ELAR Diagnostic Test are the reading-focused Text Analysis and Synthesis and the writing-focused Content Revision and Editing for Conventions. In TSIA1, each of the Reading and Writing tests has four diagnostic strands. To project the cut scores from TSIA1 to TSIA2 diagnostic strands, the theta cut scores were averaged. Equipercentile linking was performed between the TSIA1 Reading placement test and the TSIA2 ELAR reading-focused strand; and similarly between the TSIA1 Writing placement test and the TSIA2 ELAR writing-focused strand. The resulting cut scores on the theta scales are presented in Table 5.4.

Development of the Proficiency Statements

Proficiency statements for TSIA2 diagnostic strands were developed using the item mapping methodology which locates or maps each test question to a point on the score scale. This mapping helps illustrate what students know and are able to do at different score bands. Question descriptions focusing on the skills and knowledge required to respond correctly are used collectively to create proficiency statements.

Each question's position on the scale is determined by the likelihood of test takers responding to the item correctly. For TSIA2 Diagnostic Tests, questions are placed on the scale using the response probability of 0.67. Using Item Response theory (IRT), each question is placed or mapped to a score at which test takers have a 0.67 probability of selecting the correct response. Descriptions of questions mapped to a selected band corresponding to the proficiency level make up the statements of what test takers whose score fall in that range typically know and are able to do at that level.

Table 5.4:
Interim Cut Scores Through Equipercentile Linking

Test	Diagnostic Strand	Proficiency Level	TSIA1		TSIA2	
			Cut Score	% At or Above	Cut Score	% At or Above
ELAR	Text Analysis and Synthesis	Proficient	-0.913877	74.44	-0.6772	74.44
		Advanced	-0.03220	39.75	0.3975	39.75
	Content Revision and Editing for Conventions	Proficient	-0.804223	71.48	-0.5919	71.48
		Advanced	0.048438	31.99	0.6537	31.99
Mathematics	Quantitative Reasoning	Proficient	-0.62972	64.95	-0.2890	64.95
		Advanced	1.57616	0.16	1.8031	0.16
	Algebraic Reasoning	Proficient	0.52373	5.86	0.6312	5.85
		Advanced	1.46343	0.25	1.3681	0.26
	Geometry and Spatial Reasoning	Proficient	-0.11019	46.25	0.0137	46.33
		Advanced	1.56527	0.55	2.0419	0.49
	Probability and Statistical Reasoning	Proficient	-0.87637	74.11	-0.4333	74.11
		Advanced	1.62279	4.55	1.2704	4.53

5.4 Reporting

Score Reporting

In keeping with the AERA/APA/NCME Standards, TSIA2 score reports have been developed at the student and institution levels to provide their intended audiences with appropriate interpretations of the reports and guidelines outlining the appropriate use of test results. A variety of reports are available online 24/7 for all TSIA2 tests, including the following:

- Individual Score Report (ISR)
  - Generated for each test taker at the end of testing

- Shows test taker's identifying information and test scores, with conditional standard errors of measurement (CSEMs) if this option is selected by the institution
- Shows appropriated course placement if placement rules have been entered into the testing site
- If the Essay Test is taken, the ISR includes the holistic score description and dimension statements

- Essay Response Report – allows an institutional user to search and print essays submitted by test takers in response to a prompt
- Placement Roster Report – provides a list of test takers who placed into courses associated with a specific course group
- Course Roster Report – provides a list of test takers who placed into a specific course
- Score Roster Report -- customizable report that may include test takers' scores, demographic information, and answers to background questions as selected by the institution-designated user for a specific date range

Databases

TSIA2 data are stored in secure reporting databases and retained for five (5) years. All data are synched in real time with the Disaster Recovery environment so they will not be compromised during a disaster scenario. On a quarterly basis, the database system removes testing data that are more than five years old. This routine maintenance of the data stored in the system ensures that the platform provides immediate, stable, and accurate access to current student testing data.

5.5 Using Multiple Factors in Placement Decisions

Standard 12.10 of the *Standards for Educational and Psychological Testing* asserts that "In educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information." (AERA, APA, & NCME, 2014, p. 198) To provide guidance to institutions in using additional information when making placement decisions, College Board released a paper titled *Multiple Factors in College Placement Decisions,* which is available through this link: https://accuplacer.collegeboard.org/pdf/multiple-factors-college-placement-decisions.pdf.

Chapter 6 — Psychometrics

Introduction

Once the ability scores on the College Readiness Classification (CRC) tests have been estimated through the computer-adaptive test (CAT) algorithm as described in Section 3.4 of this manual, they must be transformed into scores used in reporting. This involves establishing a scale. The scale is a numerical system that conveys test performance. The first part of this chapter (Section 6.1) discusses the creation of this numerical system used to report results of the CRC tests, a procedure called scaling.

Test taker performance on diagnostic tests is not reported in scaled scores but in overall diagnostic levels 2 through 5, as well as proficiency levels (Basic, Proficient, and Advanced) for each diagnostic test strand. Determination of ability score ranges classified to each diagnostic level and proficiency levels were discussed in Chapter 5 of this manual. Results for the Essay Test are reported in raw scores.

The second part of this chapter (Section 6.2) discusses the precision or reliability of the reported results. The first portion of Section 6.2 is on the standard error of measurement of each scale score, followed by accuracy of classification in diagnostic tests. The next portion is on the interrater consistency of essay scores.

6.1 Scaling

A scale refers to a numerical system that conveys test performance. Once the ability scores (i.e., theta) on the Texas Success Initiative Assessment 2.0 (TSIA2) CRC Tests have been estimated through the CAT algorithm, they are converted to a pre-selected scale for score reporting. TSIA2 CRC Tests are computer-adaptive tests, using Item Response Theory (IRT) as the psychometric method to select questions administered to a test taker, calibrate questions, and estimate a test taker's ability. The questions chosen are based on a constantly updated evaluation of the test taker's ability after each preceding question has been answered. With IRT, the computed theta score reflects an estimate of ability that ranges from -5.0 to 5.0. Since the continuous scale of -5.0 to 5.0 is difficult to interpret for test users, an alternative scale is used to report test takers' performance on the test. For TSIA2 CRC Tests, a scale of 910 to 990 with an increment of 1 has been selected for score reporting. According to AERA/APA/NCME Standard 5.2, "the procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly" (AERA, APA, & NCME, 2014, p. 202). This section of the manual elaborates the scaling procedures for TSIA2 English language arts and reading (ELAR) and mathematics CRC tests.

Goals for the Scales

TSIA2 CRC Tests are on 81-point scales. Both ELAR and Mathematics CRC Tests have scale scores ranging from 910 to 990, with a target mean of 950 and a standard deviation of 16. These intended ranges and distributions of the scale scores are selected because they are deemed sufficient to support the

interpretation of test performance for the purpose of placing students in the appropriate college courses.

Scaling Procedures

Scaling procedures for each TSIA2 CRC Test involve creating a scale score conversion table based on the estimated student ability distribution from the CAT simulations. A mean-sigma transformation method is used to convert the continuous theta scale (i.e., range of -5.0 to 5.0) with the estimated ability distribution based on 50,000 simulees to the pre-determined scale (i.e., range of 910 to 990 with an increment of 1) with the intended distribution (i.e., mean of 950 and standard deviation of 16).

Let $SS = Scale\ Score$ with

$$\text{Range: SS} \in [910,\ 990],$$

$$\text{Mean: } \mu_{SS} = 950,$$

$$\text{Standard Deviation: } \sigma_{SS} = 16,$$

and let the first two moments of θ estimates based on the CAT simulations be $\widehat{\mu_\theta}$ and $\widehat{\sigma_\theta}$. For the ELAR CRC Test, the mean theta score $\widehat{\mu_\theta} = 0.0870$ and standard deviation of theta score $\widehat{\sigma_\theta} = 1.2260$. For the Mathematics CRC Test, the mean theta score $\widehat{\mu_\theta} = -0.3770$ and standard deviation of theta score $\widehat{\sigma_\theta} = 1.1832$.

The scaling formula is

$$SS = a + b\theta,$$

where

$$\mu_{SS} = a + b * \widehat{\mu_\theta}$$

and

$$\sigma_{SS} = b * \widehat{\sigma_\theta}.$$

From $\sigma_{SS} = b * \widehat{\sigma_\theta}$, compute the slope as

$$b = \frac{\sigma_{SS}}{\widehat{\sigma_\theta}}.$$

From $\mu_{SS} = a + b * \widehat{\mu_\theta}$, compute the intercept as

$$a = \mu_{SS} - b * \widehat{\mu_\theta}.$$

Using the estimated slope b and intercept a, a scale score is computed for each of the 50,000 test takers based on their theta score. The mean and standard deviation of the computed scale scores for the 50,000 test takers are then calculated. The values of μ_{SS} and/or σ_{SS} are adjusted and optimized as appropriate to make the computed scale score mean and standard deviation as close to 950 and 16,

respectively, as possible. Using the resulting slope and intercept based on the optimized mean and standard deviation of the scale score (shown in Table 6.1), a conversion table is constructed such that each scale score has a corresponding θ value. To satisfy a requirement of the test administration platform, -5.0 is always mapped to the lowest possible scale score 910, and 5.0 is always mapped to the max possible scale score 990.

Table 6.1:
Transformation Constants for Theta (θ) to Scaled Score Conversion for TSIA 2.0 CRC Tests

Test Name	Slope	Intercept
English Language Arts and Reading (ELAR)	13.4071	948.8015
Mathematics	14.2383	955.3075

Scale Re-centering

As discussed in the TSIA2 Standard Setting Report (Bay & Duffy, 2020), subsequent to scaling and standard setting the THECB decided that the college readiness benchmark should be mapped to the scale scores 945 and 950 for ELAR and Mathematics, respectively. This required subtracting 14 from the ELAR intercept and subtracting 13 from the Mathematics intercept. With the scale score range kept at 910 to 990, the mean and standard deviation of the scale scores intended to be 950 and 16, respectively, shifted accordingly. The mean and standard deviation of the scale scores are presented in Table 6.2.

Table 6.2:
Mean and Standard Deviation of Scale Scores

Test	Mean	Standard Deviation
English Language Arts and Reading (ELAR)	936.43	15.38
Mathematics	937.40	15.51

6.2 Reliability

What is Reliability?

Reliability is the degree to which scores arising from an assessment produce stable and consistent results. In other words, the reliability coefficient indicates the amount of consistency in scores. A score with a reliability coefficient of 1 is a perfectly reliable score, while a value of 0 means that the score is not at all reliable. If a score has a reliability of, say 0.88, one may think about it as the amount of consistency.

Reliability and Conditional Standard Error of Measurement (CSEM) of Scale Scores

The standard error of measurement (SEM) provides an estimate of the amount of error in scores. SEM is computed as

$$SEM = SD * \sqrt{1 - r}$$

where SD is the standard deviation of test scores and r is the test score reliability. Based on the SEM formula, we can see that the SEM and reliability are inversely related. That is, the more reliable the score is, the smaller standard error of measurement the score has. In contrast, the less reliable the score is, the larger standard error of measurement the score has. The SEM is especially meaningful to a test taker because it applies to a single score and uses the same units as the test.

For TSIA2 CRC Tests, a SEM is estimated for each score. That is, around each scaled score, a value is computed to indicate the level of certainty about where a test taker's true scaled score may lie, given the score that that test taker obtained.[22] Each computed value represents the variability one would expect to see in the scaled scores of a test taker of a given ability who takes the test multiple times. This is referred to as the conditional standard error of measurement (CSEM). These values may be used to report a confidence interval within which a test taker's true score might fall, given that test taker's obtained score. For example, if a test taker receives a score of 950 on a CRC test and the CSEM is 4.9, there is a 68% probability that the test taker's true score is within the 945.1 and 954.9 range. In other words, if that test taker took the test 100 times, and the range was computed each time, then approximately 68% of these ranges will contain the test taker's true score. A smaller value of CSEM indicates more precise measurement.

The CSEMs for scale scores are estimated based on simulation results after the theta to scale score conversion table is developed. The following is a summary of the procedure to estimate the scale score CSEMs:

1. Create a uniform distribution of 50,000 true thetas ranging from -5 to +5 so that all theta levels are equally represented.
2. Use the true theta distribution generated in step 1 to simulate CAT so that each true theta gets a theta estimate.
3. Convert true thetas to true scale scores using the conversion table.
4. Convert estimated thetas to estimated scale scores using the conversion table.

[22] Based on responses to all the test questions, the maximum likelihood estimate of the test taker's ability, $\hat{\theta}$, is computed as described in Section 3.4 Computer-Adaptive Test Algorithm. Using the scaling method described in Section 6.1: Scaling, this ability estimate is transformed to a scaled score. This is the test taker's observed score or the score the test taker obtained. Corresponding to the test taker's ability estimate is their true ability, θ, which is an unobservable and unmeasurable quantity we wish we could obtain, but cannot (Kingston & Stocking, 1986). The quantity θ transformed to the scaled score metric is the true scaled score.

5. At each integer true scale score point (i.e., 910 to 990), compute standard deviation of estimated scale scores as CSEM for the scale score. The unrounded scale score estimates are used in calculation for better precision.

The CSEM is an optional element for the individual student report (ISR) that is included at the discretion of the institution. It is also a very important piece of information for institutions when deciding on placement policies such as finalizing placement scores or using multiple factors in placement decisions. A new CSEM table is created whenever a CAT question pool is refreshed and is always current in the test administration platform. The most current CSEM table is available upon request from College Board.

Classification Accuracy

For test scores used for classification, the accuracy and consistency of such classifications is of interest and is considered in evaluating the quality of the tests. Recall that reliability is a measure of consistency of scores across different situations such as taking different test forms or taking the test multiple times. For a CAT, each ability estimate is effectively based on a different test form. Thus, in terms of classification consistency, one is interested in the extent to which a test taker is classified into the same category based on the multiple ability estimates when taking the test multiple times. Classification accuracy is the extent to which classification based on ability estimate is consistent with classification based on true ability. Classification accuracy is then considered an upperbound of classification consistency.

Simulation data sets generated as part of the process of developing the CRC and Diagnostic Tests were used to compute decision accuracy. In the simulated data sets, each simulated test taker has true and estimated ability (i.e., θ and $\hat{\theta}$) scores for the CRC and Diagnostic Tests. Thus, each simulee could be assigned to the appropriate categories based on the final cut scores discussed and presented in the portion of Chapter 5 covering Standard Setting and detailed in Bay and Duffy (2020). Moreover, each simulated test taker is assigned to the appropriate category based on their true ability θ, and again based on their estimated ability $\hat{\theta}$. The consistency of the classifications based on true and estimated abilities are then analyzed. For each CRC Test, there are two performance categories corresponding to whether the score is above or below the college readiness benchmark. Test takers who score below the benchmark take the Diagnostic Test. Their overall performance on the Diagnostic Test is classified into five categories: Combined Levels 1 and 2; Level 3; Level 4; Level 5; and Level 6. Additionally, test taker performance in each of the diagnostic content strands are classified into Basic, Proficient, and Advanced levels.

The overall classification accuracy was computed as the percentage of simulees who were categorized the same way based on true and estimated θs. Table 6.3 presents classification accuracy for the CRC and Diagnostic Tests. For CRC Tests, the percentage of correct classification was 90.3 for ELAR and 92.2 for Mathematics. For overall diagnostic levels, the percentage of correct classification are 80.5 and 83.1 for ELAR and Mathematics, respectively. For the diagnostic strands, the classification accuracy percentages are between 82.8 and 88.3.

Table 6.3:
Overall Classification Accuracy: Percentage of Correct Classifications

Tests and Classifications	Accuracy (%)
ELAR CRC	90.3
ELAR Diagnostic Levels	80.5
Text Analysis and Synthesis (Reading-Focused) Proficiency Levels	83.1
Sentence Structure (Writing-Focused) Proficiency Levels	82.8
Mathematics CRC	92.2
Mathematics Diagnostic Levels	83.1
Quantitative Reasoning Proficiency Levels	88.0
Algebraic Reasoning Proficiency Levels	88.3
Geometric and Spatial Reasoning Proficiency Levels	86.5
Probability and Statistical Reasoning Proficiency Levels	83.5

In addition to the overall accuracy, classification accuracy was also examined relative to each cut score. Specifically, the classification accuracy is presented by four indexes as follows:

True negative: Correct classification – students with true θs below the cut score were classified as below the cut score.
True positive: Correct classification – students with true θs above the cut score were classified as above the cut score.
False positive: Incorrect classification – students with true θs below the cut score were classified as above the cut score.
False negative: Incorrect classification – students with true θs above the cut score were classified as below the cut score.

Table 6.4 shows the four classification accuracy indexes relative to the college readiness benchmarks for ELAR and Mathematics CRC Tests. Tables 6-5 and 6-6 present classification accuracy relative to proficiency level cut scores for each ELAR and Mathematics diagnostic strands, respectively. Tables 6.7 and 6.8 report the indexes for the four cut scores delineating the five diagnostic levels of ELAR and Mathematics, respectively.

Table 6.4:
Classification Accuracy Relative to College Readiness Benchmarks

Test	Index	Percentage of Correct Classification
ELAR	True Negative	67.57
	False Positive	5.82
	False Negative	3.90
	True Positive	22.71
Mathematics	True Negative	74.64
	False Positive	4.74
	False Negative	3.07
	True Positive	17.54

Table 6.5:
Classification Accuracy Relative to Proficiency Level Cut Scores: ELAR Diagnostic Strands

ELAR Diagnostic Strands	Basic/Proficient		Proficient/Advanced	
	Index	Percentage of Correct Classification	Index	Percentage of Correct Classification
Text Analysis and Synthesis (Reading-Focused)	True Negative	22.45	True Negative	55.74
	False Positive	3.23	False Positive	5.83
	False Negative	3.65	False Negative	4.18
	True Positive	70.67	True Positive	34.25
Content Revision and Editing for Convention (Writing-Focused)	True Negative	24.52	True Negative	64.13
	False Positive	3.57	False Positive	5.45
	False Negative	4.17	False Negative	3.97
	True Positive	67.74	True Positive	26.45

Table 6.6:
Classification Accuracy Relative to Proficiency Level Cut Scores: Mathematics Diagnostic Strands

Mathematics Diagnostic Strands	Basic/Proficient		Proficient/Advanced	
	Index	Percentage of Correct Classification	Index	Percentage of Correct Classification
Quantitative Reasoning	True Negative	48.63	True Negative	95.30
	False Positive	5.28	False Positive	2.45
	False Negative	3.67	False Negative	0.62
	True Positive	42.43	True Positive	1.64
Algebraic Reasoning	True Negative	77.31	True Negative	92.01
	False Positive	5.17	False Positive	2.71
	False Negative	3.11	False Negative	1.14
	True Positive	14.41	True Positive	4.14
Geometric and Spatial Reasoning	True Negative	58.38	True Negative	96.01
	False Positive	5.91	False Positive	2.64
	False Negative	4.53	False Negative	0.43
	True Positive	31.18	True Positive	0.92
Probability and Statistical Reasoning	True Negative	43.28	True Negative	88.18
	False Positive	5.31	False Positive	5.47
	False Negative	4.18	False Negative	1.67
	True Positive	47.24	True Positive	4.69

Table 6.7:
Classification Accuracy (i.e., Percentage of Correct Classification) Relative to ELAR Diagnostic Levels

Index	Level 2/3	Level 3/4	Level 4/5	Level 5/6
True Negative	15.00	34.98	69.68	92.36
False Positive	1.62	3.30	3.71	1.82
False Negative	1.96	3.37	2.80	0.95
True Positive	81.42	58.35	23.81	4.87

Table 6.8:
Classification Accuracy (i.e., Percentage of Correct Classification) Relative to Mathematics Diagnostic Levels

Index	Level 2/3	Level ¾	Level 4/5	Level 5/6
True Negative	13.38	25.32	55.86	76.39
False Positive	1.38	1.96	2.62	2.99
False Negative	1.60	1.82	2.30	2.24
True Positive	83.64	70.9	39.22	18.38

Interrater Consistency of Essay Scores

The ELAR portion of TSIA2 has an essay component. Test takers who score above the threshold of 945 on the CAT, or those who are classified at the Diagnostic Level of 5 or 6, take the Essay portion of the ELAR test. To ensure that test takers' essays are scored reliably, an annual audit of the automated scoring is implemented for all prompts. The result of this audit is summarized here.

For the 17 Essay prompts in TSIA2, an annual national audit is conducted using the following procedure. Five percent of all essays scored by the automated scoring engine throughout the year are selected for human scoring through the Pearson Performance Scoring Center. A Stratified Sampling Method is applied to select the essays that are included in the annual audit scoring. All essays are scored by a human scorer and machine scores and human scores are analyzed for exact and adjacent agreement interrater reliability (IRR). Resolution scoring is applied to nonadjacent essays if the original IRR results are below 90% on the prompt.

During the 2020 audit, all prompts exceeded the 90% exact plus adjacent IRR agreement expectation after the initial human read. Resolution scoring by content experts would have been conducted for any prompt that fell below 90% exact plus adjacent agreement, but no prompts required resolution scoring this audit year. The value of this IRR index for each prompt is included in Table 6.9.

Another index of interrater agreement commonly used in essay scoring is the quadratic weighted kappa. It is a commonly used statistic for summarizing interrater agreement on an ordinal scale like essay scores. The value of the quadratic weighted kappa for each prompt is also included in Table 6.9.

To oversee the audit, College Board has a yearly review meeting with Pearson to go over the results of the essay score audit. Members of the Psychometrics team participate in the annual review meeting, offering feedback for potential improvements that can be applied to future audits as appropriate.

Table 6.9:
Interrater Reliability for Essay Prompts

Prompt	Number of Essays Compared	Exact plus Adjacent Agreement	Quadratic Weighted Kappa
B11- Practical Skills	618	94.3%	0.74
B17 – Success	1,565	92.8%	0.75
B24 - Acquisition of Money	1,328	95.4%	0.75
B26 - Is History Valuable	1,286	97.2%	0.79
C04 - Necessary to Make Mistakes	1,472	97.1%	0.75
C06 - Unlimited Change	1,174	96.3%	0.77
C14 - Established Rules	1,126	96.8%	0.75
C16 - Independent Ideas	1,429	95.5%	0.76
C22 - Results of Deception	714	98.2%	0.88
C24 - Books Provide Lessons	1,397	96.0%	0.83
C28 - Optimism or Realism	1,269	98.1%	0.86
C30 - Happiness Not an Accident	1,618	94.9%	0.76
D02 - DuBois Work	1,527	92.4%	0.73
D05 - Differences Among People	1,443	96.5%	0.77
D12 - Pursue External Goals	1,112	96.9%	0.77
D21 - Nontraditional Solutions	1,263	98.7%	0.85
E01 - Be Original	1,305	94.2%	0.75

Chapter 7 — Validity

Introduction

This final section covers validity. In many ways, every chapter of this manual covers validity, as all the procedures described in the previous sections attempt to ensure that Texas Success Initiative Assessment 2.0 (TSIA2) tests produce scores that are valid measures of the constructs being tested. As such, a commitment to matters of validity is of paramount importance to both College Board and the Texas Higher Education Coordinating Board (THECB) as the TSIA2 is developed, administered, and scored. A deeper examination of issues of validity are placed purposefully here at the end of this manual for the reason that by first gaining an understanding of the many processes and procedures involved in developing, administering, and scoring TSIA2, and understanding how and why those scores are interpreted for their intended uses, one can more completely comprehend the steps taken toward establishing sound validity evidence.

For TSIA2, test validity must be evaluated with respect to the degree to which test takers' scores support making appropriate placement decisions. Arguably, all the evidence presented in the previous chapters goes toward supporting this claim, from the earliest stage of test development, right up to the final interpretation of scores. The examination of validity as it relates to TSIA2 begins in the broadest terms, as Section 7.1 provides a brief overview of validity as a concept and the goals of test score validation. Then, shifting the focus to the test itself, Section 7.2 presents the evidentiary foundations behind the test content found in TSIA2 college readiness classification (CRC) and diagnostic tests. Criterion-based evidence for validity will be collected and documented when the required data from tests, decisions, and course performance are available. Specifically, a predictive placement validity study of the TSIA2 CRC tests is planned for when a full school year's worth of data has been collected.

7.1 Introduction to Validity as a Concept

Validity is not an intrinsic property of a test. Rather, it is the extent to which the inferences (interpretations) derived from test scores are justifiable from both scientific and equity perspectives. For decisions based on test scores to be valid, the use of a test for a particular purpose must be supported by theory and empirical evidence, and biases in the measurement process must be ruled out.

As many psychometricians have pointed out (e.g., Cronbach, 1971; Messick, 1989; Shepard, 1993), in judging the worth of a test, it is the inferences derived from the test scores that must be validated, not the test itself. Therefore, the specific purpose(s) for which test scores are being used must be considered when evaluating validity. For example, a test may be useful for one purpose, such as course placement, but not for another, such as college admission.

Contemporary definitions of validity in testing borrow largely from Messick, who stated that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other

modes of assessment" (Messick, 1989, p. 13). Based on this definition, validity is not something that can be established by a single study, and tests themselves cannot be labeled "valid" or "invalid." Given that validity is the most important consideration in evaluating the use of a test for a particular purpose, and such utility can never be unequivocally established, establishing that a test is appropriate for a particular purpose is an arduous task. Thus, the following facts about validity should be clear: (a) tests must be evaluated with respect to a particular purpose; (b) what needs to be validated are the inferences derived from test scores, not the test itself; (c) evaluating inferences made from test scores involves several different types of qualitative and quantitative evidence; and (d) evaluating the validity of inferences derived from test scores is not a one-time event but is a continual process.

To make the task of validating inferences derived from test scores both scientifically sound and manageable, Kane (1992, 2006) proposed an "argument-based approach to validity." In this approach, the validator builds an argument based on empirical evidence to support the use of a test for a particular purpose. Although this validation framework acknowledges that validity can never be established absolutely, it requires evidence that the test measures what it claims to measure, that the test scores display adequate reliability, and that test scores display relationships with other variables in a manner congruent with the test's predicted properties. Kane's practical perspective is consistent with the Standards (AERA, APA, & NCME, 2014), which provide detailed guidance regarding the types of evidence that should be brought forward to support the use of a test for a particular purpose. In keeping with the notion that all the statistical processes applied to a test aid in establishing validity, the 2014 Standards state that:

> *A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses.… Ultimately, the validity of an intended interpretation … relies on all the available evidence relevant to the technical quality of a testing system. [This includes] evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers.… (p. 21-22)*

To build a validity argument for a test, there are several types of evidence that can be presented. Evidence based on content involves gathering data from content experts regarding the degree to which the behaviors sampled on the test represent the behaviors the test is designed to measure. Evidence based on criterion-related information involves evaluating correlations among test scores and other variables related to the construct measured. This evidence includes predictive and concurrent as special cases that involve correlating test scores with future or current criterion performance, respectively. Other evidence for the validity of interpreting test scores involves gathering data that show test scores are indicative of the construct measured.

College Board is committed to performing a predictive placement validity study for TSIA2 CRC tests. With the notion that the CRC tests are being used by different institutions for placement to different courses, each institution is encouraged to evaluate the predictive placement validity of their placement

decisions. To support institutions in this endeavor, the College Board provides a free service through the Admitted Class Evaluation Service (ACES).

7.2 Content-Oriented Validity Evidence and Alignment

What Does TSIA2 Measure?

According to the AERA/APA/NCME *Standards for Educational and Psychological Testing*, Standard 1.11, "When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent" (AERA/APA/NCME, 2014, p. 26).

TSIA2 is intended to be used for assessing the reading, writing, and mathematical knowledge and skills essential for college and career readiness in Texas. Scores from the CRC Tests are intended to be interpreted as indicators of a student's readiness for career training programs and college, while scores from the Diagnostic Tests are intended to be used for identifying specific areas of weakness where students could benefit from additional academic support and interventions. To build an argument supporting the use of the TSIA2 CRC and Diagnostic Tests for these particular purposes, validity evidence based on test content is necessary.

This portion of the chapter describes the foundations for the decisions made about test content included in TSIA2, which is made up of the multiple-choice English Language Arts and Reading (ELAR) and Mathematics CRC and Diagnostic Tests and an Essay Test. The design and content of each test are shaped by 1) the curriculum and assessment standards[23] considered to be essential for measuring college and career readiness by Texas educators and legislators and 2) the best available research and evidence on the knowledge and skills essential for postsecondary education and career training. The first point is addressed in full in Appendix F: TSIA2 ELAR Alignments and Appendix G: TSIA2 Mathematics Alignments, which describe in detail the alignment of TSIA2 ELAR and Mathematics content to Texas's curriculum and standards, which form the backbone of TSIA2's test blueprints and content specifications. The rest of this section is devoted to the discussion of the second point—the relationship of the design and content of TSIA2 to the best available research and evidence on the knowledge and skills essential for demonstrating readiness in college and career.

Content Validity for the TSIA2 ELAR Tests

Reading-focused questions on the multiple-choice TSIA2 ELAR CRC and Diagnostic Tests assess students' comprehension and reasoning skills in relation to appropriately challenging prose passages across a range of disciplines. Writing-focused questions measure students' revision and editing skills in the context of extended prose passages as well as in single sentences across a range of disciplines.

[23] (1) Texas College and Career Readiness Standards (2018); (2) Texas Essential Knowledge and Skills (TEKS), English III (2017), Algebra II (2012); (3) AEL Content Standards 2.0; and (4) NRS EFL.

A number of noteworthy design elements strongly supported by evidence are interwoven throughout the ELAR Tests. These include:

- a focus on words in context and on effective word choice;
- the use of a specified range of text complexity aligned to college and career readiness levels of reading;
- the requirement that students work with texts across a wide range of disciplines;
- attention to source analysis and use of evidence; and
- attention to a core set of important Standard English language conventions and to effective written expression.

These key elements are described briefly below and more fully in Chapter 3: Test Development Procedures and separately in the TSIA2 ELAR test specifications (College Board, 2021a).

Words in Context. Research has shown the close link between students' vocabulary achievement and their success in reading and in school in general (Beck, McKeown, & Kucan, 2013). With a broad and deep vocabulary, readers are more likely to understand what they read and, in turn, to derive the meaning of words in the contexts in which they appear. Indeed, the role of vocabulary in reading comprehension is difficult to overstate, given the word richness of text. A quick comparison between oral and written language indicates that while the conversation of college-educated adults contains an average of 17.3 rare words per thousand, even children's books exhibit 30.9, almost double that frequency (Becker, 1977; Hayes & Ahrens, 1988; National Center for Education Statistics, 2013; National Reading Panel, 2000; Stanovich, 1986; Whipple, 1925).

Attaining skilled comprehension through vocabulary depends on how the vocabulary is acquired. Beck and her colleagues (Beck, McKeown, & Kucan, 2013) have sensibly focused on what they refer to as Tier Two words— "words that are of high utility for mature language users and are found across a variety of domains"—because they appear frequently in written texts (but uncommonly in oral language) across a wide range of subjects (p. 9). By contrast, Tier One, or basic, words require little instruction for most students fluent in English because they are generally acquired through conversation, and Tier Three words are either limited to a certain domain of knowledge—and thus are best studied as part of work in that domain—or too rare to be found with any frequency in written text. Other researchers have reached a similar conclusion about the need to concentrate instruction on high-utility words (Beck et al., 2013; Nation, 2001; Stahl & Nagy, 2006).

There is a sharp focus on vocabulary in the ELAR Tests. In the reading-focused questions, test takers are called on to determine the meaning of vocabulary in context, with an emphasis on Tier Two words and phrases. In both reading- and writing-focused questions, test takers are also presented with other vocabulary-related challenges, including analyzing word choice rhetorically and improving the precision, concision, and context appropriateness of expression.

Text Complexity. Numerous studies have highlighted the long-standing gap between the high level of challenge posed by the required readings in college-entry, credit-bearing courses and workforce training programs and the comparatively simpler readings used in much of K–12 education, including many high school courses. For example, Adams (2009), reviewing the research literature on the challenges students face reading complex texts, helped collect a range of scholarly evidence documenting a several-decades long decline in K–12 text complexity even as college and career readiness demands on students' reading skills remained high.

The ELAR Tests align the levels of text complexity represented in the tests' passages with the requirements of workforce training programs and common first-year, credit-bearing college courses. This alignment supports the emerging movement to close the preparedness gap by making text complexity a central part of the test design. Students taking the TSIA2 ELAR Tests are asked to engage with the passages selected, in part, to exhibit a range of text complexities up through and including levels comparable to those expected of students entering workforce training programs and college. To ensure that texts on the TSIA2 ELAR Tests are appropriately complex—challenging but not inaccessible to college- and career-ready students—College Board test developers make use of descriptions available in the various Texas standards, feedback from secondary and postsecondary subject matter experts, test data on student performance, and the complexity grade bands described in Appendix A: Text Complexity (Qualitative)—Reading and Writing.

Considered together, the TSIA2 ELAR CRC and Diagnostic Tests measure whether students can read, improve, and analyze texts at levels of difficulty required of incoming postsecondary students. In addition, the Diagnostic Test determines the levels of intervention or developmental education corequisites that will set students falling below the college readiness classification score on the path to success.

Disciplinary Literacy. Shanahan, Shanahan, and Misischia (2011) are prominent among those who have made the case that students' literacy development should not be seen as merely the development of generic communication skills but instead should be grounded in making students familiar with the differing literacy demands of particular fields of study. These authors claim that reading, for example, is an importantly different activity when it's done in, say, a history, a mathematics, or a chemistry context: "In addition to the 'domain knowledge' of the disciplines . . . each discipline possesses specialized genre, vocabulary, traditions of communication, and standards of quality and precision, and each requires specific kinds of reading and writing to an extent greater than has been recognized by teachers or teacher preparation programs" (Shanahan et al., 2011, p. 395).

TSIA2 ELAR Tests support a strong emphasis on disciplinary literacy through careful passage selection and question development. In both the CRC and Diagnostic Tests, test takers are expected to engage with and analyze appropriately challenging texts spanning numerous disciplines, including literature, the humanities, social science, and science, as well as texts on topics classified as practical affairs and human relationships. Moreover, while questions on the ELAR Tests do not require test takers to have prior knowledge of specific topics in various disciplines, these questions do, where possible and beneficial,

reflect differences in the ways different disciplines approach literacy. Reading-focused questions relating to a literature selection, for example, might address theme, mood, figurative language, or characterization—concepts that are generally not relevant to the sciences. Questions relating to a science selection, on the other hand, might require students to analyze research data or determine which conclusion is best supported by a study's findings—skills generally not required to comprehend literary texts.

Source Analysis and Evidence Use. Students' developed abilities to analyze source texts and, more broadly, to understand and make effective use of evidence in reading and writing are widely recognized as central to college and career readiness and success. National curriculum surveys conducted by College Board and others demonstrate that postsecondary instructors rate high in importance such capacities as summarizing a text's central argument or main idea, identifying rhetorical strategies used in a text, and recognizing logical flaws in an author's argument, as well as writing analyses and evaluations of texts, using supporting details and examples, and developing a logical argument (Achieve, Inc., The Education Trust, & Thomas B. Fordham Foundation, 2004; ACT, Inc., 2009; College Board, 2019; Kim, Wiley, & Packman, 2012; Seburn, Frain, & Conley, 2013).

TSIA2 ELAR Tests support an emphasis on source analysis and evidence use throughout the assessments. Reading-focused questions ask test takers to answer based on what is stated and implied in texts across a range of disciplines. Many writing-focused questions ask test takers to develop, support, and refine claims and ideas in multiparagraph passages and to add, revise, or delete information in accordance with rhetorical purpose.

Language Conventions and Effective Language Use. In addition to vocabulary knowledge and use, skilled expression in language includes understanding and observing the conventions of Standard English and, more generally, making informed, thoughtful grammatical choices. Knowledge of conventions includes learning and adhering to language "rules" governing conventional expression, as well as knowledge of the practices that lend precision and clarity to writing, aid comprehension, and facilitate academic success. Grammatical choices represent the relationships between writers and their world and express how writers attend to the words of others and position themselves in relation to others (Micciche, 2004).

The writing-focused questions on the ELAR Tests support a thoughtful emphasis on language use and language conventions. Students are assessed in the context of high-quality multiparagraph passages that must be revised and edited as well as single sentences that must be completed appropriately. Language conventions and effective language use are also emphasized in reading-focused questions that address students' capacity to analyze word choice rhetorically.

Content Validity for the TSIA2 Essay Test

The TSIA2 Essay Test is designed to help postsecondary administrators assess students' readiness to successfully meet the writing demands of introductory credit-bearing courses. It assesses students' ability to successfully write an original essay in which they develop a point of view or position in

response to a writing task using reasoning, personal experience, observations, and an appropriate rhetorical approach. Compared to multiple-choice questions, this direct writing assessment may offer opportunities for students to present a more comprehensive picture of what they know and are able to do. For the Essay Test to yield meaningful outcomes, and because test scores are used to make important decisions that affect students' path to college, it is critical that

- each writing prompt successfully elicit a performance of writing;
- the scoring rubric clearly articulates the definitions of the construct measured (i.e., writing ability in the context of college-bound students);
- the Intelligent Essay Assessor (IEA) that electronically scores the essays be trained to interpret these scoring criteria.

The next section discusses the first two imperatives above, while information on scoring and IEA is available in Chapter 5: Interpretation and Application of Results.

Research suggests that some variables unrelated to the focal measurement construct (e.g., unnecessary linguistic complexity, cultural biases in construction of prompts) can affect the trustworthiness of test scores (Abedi, 2006; Solano-Flores & Trumbull, 2003; Solano-Flores, 2008), thus negatively impacting an assessment's usefulness as a tool for evaluating student learning and informing instruction. To make sure the writing tasks (i.e., prompts) and criteria for evaluating responses are relevant to the construct and intended score interpretation, test developers work closely with writing experts, including writing faculty members from high schools, two-year colleges, and four-year colleges from around the country, to define and operationalize the construct, review and refine prompts, and review scoring criteria and field test data. Additionally, an audit of IEA scoring is undertaken annually. The audit process and results are discussed in Chapter 6: Psychometrics.

TSIA2 Essay Test prompts must elicit the components of writing that are included in the definition of the construct assessed—no more and no less, as to do the former would result in construct-irrelevant variance and to do the latter would lead to underrepresentation of the construct. To support this imperative, two important design elements are interwoven into the Essay Test. These include:

- accessible prompts that allow test takers to respond in a variety of ways; and
- a focus on core writing skills required of most entry-level, credit-bearing college courses.

These key elements are discussed briefly below and in more detail in Chapter 3: Test Development Procedures.

Accessible Prompts. Because all test takers should have an opportunity to demonstrate their knowledge and skills in the construct being assessed, TSIA2 Essay prompts emphasize accessibility in relation to the writing task. Prompts are designed to stimulate critical thinking and to allow test takers to fulfill the writing task by drawing on a wide range of ideas and personal experiences. Prior topic-specific knowledge, including knowledge of U.S. culture or norms, isn't tested. Prompts are also relatively short

(up to 80 words). Issues and topics posed in the prompts are relevant to any number of fields and are stated in a straightforward manner, while challenging language and vocabulary, and complex structures are minimized.

Skills That Matter Most. The Essay Test focuses on a core set of skills students need in order to be proficient writers. The six essential skills assessed are purpose and focus (the extent to which the writer presents information in a unified and coherent manner, clearly addressing the issue); organization and structure (the extent to which the writer orders and connects ideas); development and support (the extent to which the writer develops and supports ideas); sentence variety and style (the extent to which the writer crafts sentences and paragraphs demonstrating control of vocabulary, voice, and structure); mechanical conventions (the extent to which the writer expresses ideas using Standard English conventions); and critical thinking (the extent to which the writer communicates a point of view and demonstrates reasoned relationships among ideas).

Content Validity for the TSIA2 Mathematics Tests

The overall aim of the TSIA2 Mathematics Tests is to assess students' fluency with, understanding of, and ability to apply the mathematical concepts, skills, and practices that are most strongly prerequisite for and useful across a range of college majors and careers.

As with ELAR Tests, a number of noteworthy design elements strongly supported by evidence are interwoven through the Mathematics Tests. These include:

- a focus on content that matters most for college and career readiness;
- an emphasis on problem solving and data analysis; and
- the inclusion of both calculator and no-calculator questions as well as attention to the use of a calculator as a tool.

These key elements are discussed briefly below and more fully in Section 3.1: Guiding Principles of College Board's Test Development Process of Chapter 3: Test Development Procedures and separately in the TSIA2 Mathematics test specifications (College Board, 2021b).

Focusing on Content That Matters Most. Across the country, evidence suggests a possible disconnect in mathematics between the K–12 and higher education systems. In one national survey, high school teachers and postsecondary instructors were asked whether students were leaving high school very well prepared for college-level mathematics. While 37 percent of high school teachers said yes, only 4 percent of postsecondary instructors agreed (Sanoff, 2006). Surveys of postsecondary faculty and studies of entry-level postsecondary course demands have repeatedly pointed to the conclusion that postsecondary instructors value greater command of a smaller set of prerequisites over shallow exposure to a wide array of topics (ACT, Inc., 2009).

In October 2013, the Council of Chief State School Officers released a set of summative assessment principles for ELA/literacy and mathematics assessments aligned to college and career readiness

standards. These assessment principles are meant to form the basis for states' evaluations of their assessment systems. The principles greatly stress the importance of focusing summative assessments on what matters most. The very first alignment principle in mathematics is that of "focusing strongly on the content most needed for success in later mathematics." As the document notes, "In a [college- and career-ready] aligned assessment system . . . high school focuses on widely applicable prerequisites for careers and postsecondary education" (2013, p. 2).

One of the most important ways the TSIA2 Mathematics Tests address the gap between postsecondary and K–12 expectations is through their concentrated focus on the content that matters most for postsecondary education. In their report on the results of a national survey, Conley et al. (2011) reinforced the conclusion that some areas of mathematics require much stronger emphasis than others. As seen in Conley's data, the importance of algebra is unmistakable, while other mathematics topics have a more mixed profile, typically including more material that isn't as relevant to and/or prerequisite for most postsecondary work. The data from this study directly support the content choices made in the Mathematics Tests.

Problem-Solving and Data Analysis. There is ample evidence that problem-solving and data analysis—the abilities to create a representation of a problem, consider the units involved, attend to the meaning of quantities, and know and use different properties of operations and objects—are important for college and career readiness and for life more generally. Quantitative literacy is part of participation in a democracy; it's important to employers, who need students who can use mathematics outside of the classroom; and it's important not only for science, technology, engineering, and mathematics (STEM) fields but also for a wide range of college majors (Conley, 2006; Conley, McGaughy, Brown, van der Valk, & Young, 2009; National Council on Education and the Disciplines, 2001).

One study by the National Center on Education and the Economy (2013) that analyzed the actual mathematical demands of course syllabi and assignments in two-year institutions also supports the emphasis of the TSIA2 Mathematics Tests on problem-solving and data analysis. The study found that students pursuing two-year degree programs must be able to work with multistep problems involving ratios, proportional relationships, percentages, unit conversions, and complex measurement problems.

Such problems are an ideal connection point for science and for college and career readiness because so many of the quantities in applied science involve proportional relationships and/or are formed by division (such as rates and densities). In addition, the Probabilistic and Statistical Reasoning questions on the Mathematics Tests contain some multistep problems that may require students to analyze data from a graphical representation and break down the question into multiple steps in order to solve it.

7.3 Final Remarks on Validity of TSIA2

As part of College Board's contractual obligation to the THECB, the predictive placement validity of each of the TSIA2 tests will be investigated. The validity studies to be conducted will investigate the relationship between performance on the tests and success on the introductory credit-bearing college

courses for which the tests are used for placement. The timing of the studies and data used will be carefully planned, with input from the THECB. Data from a full school year of test administration is often sufficient. Recognizing that TSIA2 was launched in the middle of the COVID-19 pandemic, prudence would suggest judiciousness with regard to the data that will be considered sufficient so that study results will be helpful, generalizable, and meaningful.

References

Abedi, J. (2006). Language issues in item-development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Erlbaum.

Achieve, Inc., The Education Trust, and Thomas B. Fordham Foundation. (2004). *The American Diploma Project: Ready or not: Creating a high school diploma that counts*. Washington, DC: Achieve, Inc.

ACT, Inc. (2007). *ACT national curriculum survey 2005–2006*. Iowa City, IA: Author.

ACT, Inc. (2009). *ACT national curriculum survey 2009*. Iowa City, IA: Author.

Adams, M.J. (2009). The challenge of advanced texts: The interdependence of reading and learning. In E. H. Hiebert (Ed.). *Reading more, reading better: Are American students reading enough of the right stuff?* New York: Guilford.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Monticello, NY: Marcel Dekker, Inc.

Bay, L., & Duffy, L. (2020). *Texas Success Initiative Assessment 2.0 standard setting report: An online implementation.* College Board: New York.

Beck, I.L., McKeown, M.G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd Ed). New York: Guilford Press.

Becker, W.C. (1977). Teaching reading and language to the disadvantaged — What we have learned from field research. *Harvard Educational Review*. *47 (4),* 518-543.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Cai, L. (2017). FlexMIRT: Flexible multilevel multidimensional item analysis and test scoring (version 3.51) [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Camilli, G., & Shepard, L.A. (1993). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Chuah, S. C., Hare, D., Bay, L., & Proctor, T. (2020). Automated test assembly: Case studies in classical test theory and item response theory. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment*. Charlotte, NC: Information Age Publishing INC.

Cizek, G. J. (Ed.) (2012). *Setting performance standards: Foundations, methods, and innovations.* New York: Routledge.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17(1), 31-44.*

College Board (2017). *Texas Success Initiative Assessment technical manual*, New York, NY: College Board.

College Board. (2018). ACCUPLACER© *program manual.* New York, NY: College Board. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/accuplacer/accuplacer-program-manual.pdf

College Board. (2019). *College Board national curriculum survey.* New York: NY: College Board. Retrieved from: https://collegereadiness.collegeboard.org/pdf/national-curriculum-survey-report.pdf

College Board. (2020a). *Texas Success Initiative Assessment 2.0 administrator's manual*. New York, NY: College Board.

College Board. (2020b). *Texas Success Initiative Assessment 2.0 technical manual*, New York. NY: College Board.

College Board. (2020c). *Texas Success Initiative Assessment 2.0 student informational brochure*. New York, NY: College Board.

College Board. (2020d). *Texas Success Initiative Assessment 2.0 interpreting your scores*. New York, NY: College Board.

College Board. (2021a*). TSIA2 English language arts and reading test specifications* (Version 1.4). New York, NY: College Board.

College Board. (2021b*). TSIA2 mathematics test specifications* (Version 1.4). New York, NY: College Board.

Conley, D.T. (2006). *College Board Advanced Placement® best practices course study*. Eugene, OR: Educational Policy Improvement Center.

Conley, D.T., Drummond, K.V., de Gonzalez, A., Rooseboom, J., & Stout, O. (2011). *Reaching the goal: The applicability and importance of the Common Core State Standards to college and career readiness*. Eugene, OR: Educational Policy Improvement Center.

Conley, D.T., McGaughy, C., Brown, D., van der Valk, A., & Young, B. (2009). *Validation study III: Alignment of the Texas College and Career Readiness Standards with courses in two career pathways*. Eugene, OR: Educational Policy Improvement Center.

Council of Chief State School Officers. (2013). *States' commitment to high-quality assessments aligned to college- and career-readiness*. Washington, DC: Author.

Cox, D.R. and Snell, E.J. (1989). *Analysis of binary data*. (2nd ed.). Chapman & Hall.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike & Angoff, W.G. (Eds.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Fan, M. (2007). *A comparison of automated test assembly programs for constructing parallel test forms-new and old*. Paper presented at the annual meeting of the Psychometric Society, Tokyo, Japan.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hayes, D.P., & Ahrens, M.G. (1988). Vocabulary simplification for children: A special case of 'Motherese'? *Journal of Child Language. 15,* 395–410.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale NJ: Erlbaum.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329–349.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kim, Y., Wiley, A., & Packman, S. (2012). *National curriculum survey on English and mathematics*. New York: College Board.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

Kingston, N. M., & Stocking, M. L. (1986). *Psychometric issues in IRT-based test construction.* Paper presented at the annual meeting of the American Psychological Association (Washington, DC, August 22-26, 1986).

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag, p. 162.

Lewis, D.M., Green, D.R., Mitzel, H.C., Baum, K., & Patz, R.J. (1999). *The bookmark standard setting procedure: Methodology and recent implications*. Monterey, CA: McGraw-Hill.

Lord F.M. (1980) *Applications of Item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McGaughy, C., Bryck, R., & de González, A. (2012). *California Diploma Project technical report III: Validity study; validity study of the health sciences and medical technology standards.* Eugene, OR: Educational Policy Improvement Center.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed*., pp. 13–103). New York: American Council on Education and Macmillan.

Micciche, L. R. (2004). Making a case for rhetorical grammar. *College Composition and Communication. 55*. 10.2307/4140668.

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

Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York: Cambridge University Press.

National Center for Education Statistics. (2013). *The nation's report card: Vocabulary results from the 2009 and 2011 NAEP reading assessments* (NCES 2013–452). Washington, DC: Institution of Education Sciences, U.S. Department of Education.

National Center on Education and the Economy. (2013). *What does it really mean to be college and work ready? The mathematics required of first year community college students*. Washington, DC: Author.

National Council on Education and the Disciplines. (2001) *Mathematics and democracy: The case for quantitative literacy*. Princeton, NJ: Author.

National Reading Panel (U.S), National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read; An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction; Reports of the subgroups.* Washington, DC: National Institute of Child Health and Human Development.

National Reporting System for Adult Education. (2021). *Technical assistance guide for performance accountability under the workforce innovation and opportunity act*. Washington, DC: Division of Adult Education and Literacy Office of Career, Technical, and Adult Education U.S. Department of Education Retrieved from https://nrsweb.org/sites/default/files/NRS-TA-Mar2021-508.pdf

Sanoff, A.P (2006). What professors and teachers think: A perception gap over students' preparation. *Chronicle of Higher Education*. *Vol 52, Issue 27*, p. B9.

Seburn, M., Frain, S., and Conley, D.T. (2013). *Job Training Programs Curriculum Study*. Eugene, OR: Educational Policy Improvement Center.

Segall, D. O., & Davey, T. (1995). *Some new methods for content balancing adaptive test*. Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

Shanahan, C., Shanahan, T., and Misischia, C. (2011). Analysis of expert readers in three disciplines: History, mathematics, and chemistry. *Journal of Literacy Research*. *43 (4),* 393-429.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Sireci, S. G. (2001). *Analysis of differential item functioning across females and males on the levels of English proficiency tests*. Report prepared for College Board.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, *37(4),* 189–199.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32(2),* 3–13.

Stahl, S.A., & Nagy, W.E. (2006). *Teaching word meanings*. Mahwah, NJ: Erlbaum.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly. 21*, 360-406.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Texas Education Agency. (n.d.). *Texas essential knowledge and skills for English language arts and reading; Subchapter C. high School*. Retrieved from http://ritter.tea.state.tx.us/rules/tac/chapter110/ch110c.html

Texas Education Agency. (n.d.). *Texas essential knowledge and skills for mathematics; Subchapter C. high school*. Retrieved from http://ritter.tea.state.tx.us/rules/tac/chapter111/ch111c.html

Texas Higher Education Coordinating Board. (n.d.). *Revised college and career readiness standards for English/Language arts July 2018*. Report Center. Retrieved from https://reportcenter.highered.texas.gov/agency-publication/miscellaneous/revised-ela-standards/

Texas Higher Education Coordinating Board. (n.d.). *Revised college and career readiness standards for mathematics July 2018*. Report Center. Retrieved from: https://reportcenter.highered.texas.gov/agency-publication/miscellaneous/revised-math-standards/

Texas Workforce Commission. (n.d.). *Texas adult education and literary content standards 2.0. Adult education*. Retrieved from https://www.twc.texas.gov/students/adult-education

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* Boston, MA: Kluwer Academic Publishers.

Wainer, H. (2000). *Computerized-adaptive testing: A primer (2nd ed).* Mahwah, NJ: Erlbaum.

Whipple, G.M. (1925). *Report of the National Committee on Reading: Twenty-fourth yearbook of the National Society for the Study of Education, Part 1*. Bloomington, IN: Public School Publishing Company.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human ResourcesResearch and Evaluation, Department of National Defense.