Critical Evidence 2.1.1

SAT Suite Technical Manual (October 2017)

CollegeBoard SAT®

SAT® Suite of Assessments Technical Manual

CHARACTERISTICS OF THE SAT

SAT® Suite of Assessments Technical Manual

CHARACTERISTICS OF THE SAT

October 2017

Contents

Foreword

In 2013, the College Board embarked on one of its largest initiatives ever: the redesign of its flagship program, the SAT, and its related assessment, the PSAT/NMSQT. This redesign effort was driven by the lack of progress we've seen in the last decade to increase the college readiness rates of our high school graduates, which demanded that we do something different. Simply redesigning the SAT would not be sufficient. So we, as an organization, took on the challenge of doing more than simply designing a Suite of Assessments that would report test scores. We determined that we would also deliver opportunities to middle school and high school students, to help them successfully navigate the path to college.

In 2015-16, we introduced the redesigned SAT and three PSAT-related assessments: PSAT/NMSQT, PSAT 10, and PSAT 8/9, comprising the SAT Suite of Assessments. Not only do these assessments measure the skills and knowledge needed for college readiness at grade-appropriate levels, but each of these assessments is also scaled to a single, common metric. The result is a longitudinal assessment system, beginning in grade 8 and continuing through grade 12, that helps students and teachers monitor progress toward college readiness and can start early in the preparation process, when effective interventions can be made.

The SAT Suite connects students to opportunities to help them navigate the path to college, including: deep practice of college readiness skills through our partnership with Khan Academy (available to all students free of charge); an expanding array of college scholarships that make it possible for more students to defray the rising costs of higher education; an increase in the number and kinds of fee waivers for income-eligible students, not only for the assessments but for college applications; and educational and career planning tools to help students explore their interests, identify possible college majors based on those interests, and choose colleges to which they want to apply.

It's important to recognize that we could only demand more of assessment if the SAT Suite proved to be technically sound; effectively serving its stated purposes and providing well-documented evidence of its psychometric properties.

This Technical Manual documents the processes and outcomes of the design of the SAT Suite of Assessments. More importantly, this Manual represents a baseline of evidence supporting the test development and psychometric quality of the Suite. The Manual will be supplemented each year as the Suite is administered to more and more students nationally and internationally. Research-based evidence is the hallmark of College Board's work, and we will continue to evaluate and refine our assessments and how they can be effectively used to promote student readiness based on the results of our ongoing research.

I am proud of the hard work that went into the design of the SAT Suite of Assessments and into the development of this Technical Manual. The many College Board team members who contributed to this Manual were committed to presenting this documentation in an easy-to-read format, with clear, concise evidence supporting the stated uses of the assessments. Ultimately, those who interpret and use the SAT Suite will evaluate whether we achieved these goals. We welcome any and all suggestions for improvement as we continue to update this Manual in the future.

Cynthia Board Schmeiser
Special Advisor to the President
The College Board

Contributors

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

Chapter Leads

Hui Deng, Emily Shaw, Jane Dapkus, Andrew Courchane, Sherral Miller, Cynthia Schmeiser

Writers/Reviewers

Maureen Ewing, Tim Moses, YoungKoung Kim, Judit Antal, Amy Hendrickson, Pamela Kaliski, Michael Chajewski, Rosemary Reshetar, Michael Walker, Weiwei Cui, Joseph Grochowalski, Burcu Kaniskan, Anita Rawls, Xiuyuan Zhang, Lei Wan, Chuah Siang Chee, Thomas Proctor, Prinyank Patel, Nikhil Pargaonkar, Jay Happel, Paula Cunningham, Carly Bonar, Jim Patterson, Daming Zhu, Nancy Burkholder, Rosa Baek, Martha Bell, Jennifer Karan, Aaron Lemon-Strauss, Martha Morris, Stephanie Morrison, Jennifer Merriman, Kelly Godfrey, Jessica Marini, Jeffrey Wyatt, Sanja Jagesic, Betsey Walters, Suzette Stone Busa

Technical Manual Working Group

Mark Syp, Chief Editor

Tim Moses, Chief Psychometrician

Oliver Zhang, Group Lead

Gail Mitnik, Project Manager

Karin O'Connor, Quality Control Support

Len Carmichael, Quality Control Support

Leadership Reviewers

Kevin Sweeney, Psychometrics

Sherral Miller, Assessment Design and Development

Jane Dapkus, College Readiness Assessments

Cynthia Schmeiser, Office of the President

Rosemary Reshetar, Psychometrics

Special Thanks

Jack Buckley, Gerald Melican, Carol Whang, Kristopher John

Preface

Purpose of Manual

The purpose of this technical manual is to provide higher education, K–12 educators, students, and any others who use or who are interested in using the SAT® Suite of Assessments with information about the technical qualities of the SAT Suite. This manual contains information pertaining to the purpose of the assessments and the rationale and principles behind the SAT Suite. It also includes the content of the assessments; the procedures and processes that are undertaken in the creation, administration, and scoring of the assessments; how to interpret SAT Suite scores; the accuracy of the scores from a measurement perspective; and evidence that bears on the validity of interpretations made on the basis of the scores.

The 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, particularly in light of the recent SAT Suite redesign, the operational data for which will be reported over the next few years. To this end, this manual was conceived as a "living document," and in order to keep the information in the manual as current as possible for users (Standard 7.14), it will be updated as more information becomes available.

It is important to note that the information in this manual pertains to the redesigned SAT that was first administered in March 2016. For our purposes, this will be the test we are referring to when we use the term "SAT." When necessary to draw a distinction between this iteration of the test and that administered prior to March 2016, we will use the terms "new SAT" and "old SAT," respectively. Similarly, when we use the terms "PSAT/NMSQT®," "PSAT™ 10," or "PSAT™ 8/9," we are referring to the tests first administered in fall 2015. When necessary to draw a distinction between this iteration and prior ones, we will use "new" and "old" the same way they are used when discussing the SAT.

Manual Contents

For ease of reading and understanding, this manual is structured in a manner matching that of "the lifecycle of the test." It provides insights about the SAT Suite, from our earliest conceptions of the assessment's design, all the way through test development, administration, scoring, and the interpretation of those scores for intended uses.

As its name implies, Chapter 1: Overview provides an overview of the content of the SAT Suite of Assessments and a discussion of the rationales and guiding principles behind the redesigned assessments. These guiding principles include fairness, reliability, and validity. Chapter 2: Fairness provides an examination of fairness as it relates to the SAT Suite, given the crucial role it plays in all stages of test design and administration.

Chapter 3: Test Development Procedures moves from the principles guiding the design to the processes used to design the assessments. This chapter provides a special focus on how these efforts are essential toward creating assessments that produce scores that are valid for their intended uses. After detailing the creation of the assessments in the test development section, Chapter 4: Testing Requirements describes the procedures used to administer the tests, including test security measures, in a manner that supports fair and valid uses of the tests.

With the assessments having been administered, we then turn our attention toward the scores that are produced. Chapter 5: Interpretation and Application of Results looks at the scoring procedures and analyses used to ensure scores that are valid, reliable, and fair for intended uses. Chapter 6: Psychometrics takes this one step further, as it demonstrates the ways in which the College Board evaluates the scores that come from our assessments.

Chapter 7: Validity devotes an entire chapter to the guiding principle of validity. As this manual hopefully makes apparent, validity considerations permeate every aspect of the SAT Suite of Assessments and, in theory, can be discussed in nearly every chapter of the manual. We have chosen to address validity at this point, as it represents in a significant way the culmination of our efforts. Validity evidence takes the results of all of our previous analyses and addresses whether the assessments can be used to determine college readiness and success, the overall goal of the SAT Suite.

CHAPTER 1

Overview

This section provides an overview of the SAT Suite of Assessments. It offers insight into not only the content of the assessments in the SAT Suite but also our motivations for the recent redesign and the fundamental guiding principles behind the design of the SAT Suite.

Section 1.1 takes an initial look at the SAT Suite of Assessments. In addition to providing an overview of the features of the SAT and the PSAT-related assessments, it discusses the purpose and several of the intended uses of the assessments. The section also includes a brief primer on the concepts of validity, reliability, and fairness, with a specific focus on their relevance to the SAT Suite.

Section 1.2 puts the SAT Suite into a broader context by providing a brief history of the development of the SAT and the PSAT-related assessments, including the redesign. The section offers insight into the rationale behind the redesign and the foundational tenets of the assessment. It also details the benefits the SAT Suite brings to students and parents, admission officers, and K–12 educators. This includes a special focus on the role that student preparation, through challenging high school coursework or a more formal series of test preparation, has on SAT outcomes.

Section 1.3 is a description of the content of the SAT and PSAT-related assessments. We demonstrate how the principles and tenets discussed in the previous sections were applied to create assessments that provide a sound evaluation of student achievement, the best work of the classroom, and those topics that are the most important to college and career readiness.

1.1 Introduction

Brief Description of the SAT

The SAT, the College Board's flagship college and career readiness assessment, is a key component in the SAT Suite of Assessments, which contains the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9 as grade-appropriate assessment options for middle school and high school students.

For nearly a century, the SAT has been used successfully worldwide, in combination with factors such as high school grade point average (HSGPA), to assess student preparedness for and to predict student success in postsecondary education. In the graduating class of 2017, 1.8 million test takers took the SAT, the results of which were used by thousands of high school educators and postsecondary admission officers around the world (College Board, 2016a). Since its launch in 1926, the SAT has been used to help millions of students connect with college success. The College Board's goal is to ensure that all students we serve have access to resources that can help them prepare for, and make, a successful transition to college.

For the last decade, fewer than half of all SAT test takers graduated from high school academically prepared for the challenges of college-level coursework. This number has remained virtually unchanged over the last several years. In response to this growing need, we have committed to an opportunity agenda focused on propelling high school students

into the opportunities they have earned. The redesign of the SAT and the creation of the SAT Suite are major components of this agenda.

The SAT Suite of Assessments was introduced as part of the College Board Readiness and Success System, a system designed to make it easier for students to navigate a path through high school, college, and career. The SAT Suite comprises the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9, all of which now focus comprehensively on the few durable skills that evidence shows matter the most for college and career success.

Features of the SAT and PSAT-Related Assessments

The SAT is composed of an Evidence-Based Reading and Writing section (which includes a Reading Test and a Writing and Language Test), a Math section, and an optional Essay. Test takers have three hours (plus an additional 50 minutes for the optional essay) to complete the SAT, and less than three hours for each of the PSAT-related assessments. (See the Content Specifications in Appendix 1: Overview for an exact breakdown of the times given for each assessment.)

The SAT features a continued emphasis on reasoning, alongside a clearer, stronger focus on the knowledge, skills, and understandings that are the most important for college and career readiness and success. It also places a greater emphasis on the meaning of words in extended contexts and on how word choice shapes meaning, tone, and impact. Another feature of the assessment is rights-only scoring (a point for a correct answer but no deduction for an incorrect answer; blank responses have no impact on scores). The PSAT-related assessments have been designed to measure the same domains as the SAT, but at grade-appropriate levels.

The SAT Essay is optional and is given at the end of the SAT. It is at the discretion of postsecondary institutions as to whether they require the SAT Essay for admission. Test takers have 50 minutes to produce a written analysis of a provided source text. The SAT Essay is designed to test reading, analysis, and writing skills.

The score reporting for the SAT is on a scale ranging from 400 to 1600, with a scale ranging from 200 to 800 for Evidence-Based Reading and Writing and from 200 to 800 for Math. Essay results are reported separately and have a scale ranging from 2 to 8 on each of the three performance areas. For every test, the subscores provide added insight into student achievement for students, parents, admission officers, educators, and counselors. The scale range for each assessment, including PSAT-related assessments, can be found in Appendix 1: Overview.

Statement of Purpose (Intentions and Uses of the SAT Suite of Assessments)

The primary purpose of the SAT Suite of Assessments is to determine the degree to which students are prepared to succeed, both in college and in workforce training programs. All assessment content, which was developed using the current research identifying the knowledge and skills most essential to college/career readiness and success, align with this purpose. Each test within the SAT 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 training program readiness. Because the SAT Suite assesses the content

that research shows matters the most for college and career readiness, the resulting scores provide meaningful information about a student's likelihood of succeeding in college. With this being said, the SAT Suite results shouldn't be used as the sole source of information for high-stakes decisions.

The SAT Suite provides data that are used for many purposes by different users. The three key users are higher education, K–12 educators, and students. In keeping with best practices and AERA/APA/NCME Standards, the SAT Suite's primary intended uses and interpretations for each group of primary users are discussed in the following paragraphs, with a rationale presented for each use. A summary of the evidence and theory bearing on each intended interpretation is presented in Chapter 7: Validity (AERA/APA/NCME, 2014).

Intended Uses and Interpretations

Evaluating and monitoring students' college and career readiness (For use by K–12 educators and students). The SAT College and Career Readiness Benchmarks (SAT benchmarks) serve as a challenging, meaningful, and actionable performance indicator for college and career readiness of students. States, districts, and schools use the SAT benchmarks to monitor and determine what proportion of their student body has a high likelihood of success in college-entry coursework. With the redesigned assessments, benchmark information is also provided to individual students. The SAT benchmark is not intended for high-stakes decisions such as restricting student access to challenging coursework or discouraging aspirations of attaining higher education. For more information on the SAT benchmarks, see Section 7.5. Grade-level benchmarks are also provided through the PSAT-related assessments. The grade-level benchmarks indicate whether students are on track for college and career readiness. They are based on expected student growth toward the SAT benchmarks at each grade.

Making college admission and college course placement decisions (For use by higher education). The SAT is intended for use in college admission and course placement decisions. The SAT provides rich information on a student's level of preparedness for college-level work that helps admission professionals to make more informed selection decisions. Over time, colleges and universities will be able to use the more detailed SAT score information, along with other data, to make more refined course placements for their students. This use is supported by predictive validity evidence examining the prediction of postsecondary outcomes as well as the accuracy of course placement decisions. See Chapter 7: Validity for more information on validity as it pertains to the SAT.[1]

Monitoring student progress through a vertically scaled suite of assessments (For use by K–12 educators, students). Every test in the SAT Suite is reported on a common vertical scale, with the SAT as the capstone measure (see Section 7.3 for vertical scaling information). The SAT scales are established on a nationally representative college-bound population of juniors and seniors, and the scales for the PSAT/NMSQT and PSAT 10 tests and PSAT 8/9 test are established on a nationally representative sample of students in grades 10

[1] When College Board tests are used for admission purposes, the responsible officials and selection committee members should use SAT scores in conjunction with other indicators, such as the secondary school record (grades and courses), interviews, personal statements, writing samples, portfolios, recommendations, etc., in evaluating the applicant's admissibility at a particular institution (College Board, 2011).

and 9, respectively. Establishing the scales in this manner allows for appropriate inferences of student growth and progress toward being on track for college and career readiness from year to year prior to taking the SAT. One is then able to make statements about a student's level of preparedness for college and career based on SAT performance. Students can track their own progress by using score information to identify instructional areas needing improvement and then engage in practice opportunities that will help them become more prepared for college-level work.

Contributing to high school course placement decisions (For use by K–12 educators, students). All assessments across the SAT Suite provide information about a student's readiness for particular AP courses.[2] AP Potential results provide a more challenging indication of college readiness in a particular subject through actual student performance on the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9, and the AP Exams. These results can provide students with information about what college-level classes they are ready for in high school and courses for which they need to seek additional supports before enrolling.

Contributing to scholarship decisions (For use by higher education). Colleges and organizations that award scholarships, loans, and other types of financial aid to students may tie such assistance to students' academic qualifications, as reported by SAT scores.

SAT Suite scores shouldn't be used as the single measure to rank or rate teachers, educational institutions, districts, or states. Users should exercise care when attempting to interpret test results for a purpose other than the intended purposes described in this chapter. The College Board isn't aware of any compelling validation evidence to support the use of any of the SAT Suite of Assessments, or other educational achievement measures, as the principal source of evidence for teacher or school leader evaluation. Assessment data, when properly used and subjected to several constraints, can be used *in conjunction with other educational outcome measures* to make inferences about school quality and educational quality, including teaching and learning. For further examples of uses of College Board test scores that should be avoided, see Appendix B of the *Guidelines on the Uses of College Board Test Scores and Related Data*, available online (College Board, 2011).

Overview of Fairness, Reliability, and Validity and Their Relevance to The SAT Suite

As the utility of the SAT Suite is so intricately tied to the evidence available on the fairness, reliability, and validity of the associated scores, this manual presents all relevant evidence in these areas to date. This information is interwoven throughout many of the following chapters, and the manual also has dedicated sections on these topics. The College Board is committed to adhering to the guidelines and standards outlined in the AERA/APA/NCME *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014), which is the preeminent publication in the field of testing that outlines the criteria for test development as well as the evaluation of tests, testing practices, and the interpretation of test scores for their intended uses. As such, this technical manual explains and provides information

[2] AP Potential uses scores from the SAT, PSAT/NMSQT, and PSAT 10 to provide predictions for 23 AP Exams. Scores from the PSAT 8/9 are used to identify ninth graders with potential to succeed in AP World History and AP European History, the two AP courses offered most often to 10th graders. Eighth graders will not receive AP Potential results directly.

on fairness, reliability, validity, and other critical topics that test users need to interpret and evaluate the technical quality of the assessments in the SAT Suite.

Fairness in testing touches on critical issues such as the equitable treatment of all test takers during the testing process, the lack or absence of measurement bias in test scores, and test taker access to the material that is being assessed, as well as ensuring that the interpretations of individual test scores for intended uses are valid (AERA/APA/NCME, 2014). Fairness is also addressed in test-equating procedures that are used to control for unintended difficulty differences and to produce scores that are interchangeable across different test forms. Information on the fairness of the SAT Suite can be found in essentially all chapters of this manual, in order to highlight its central importance and foundational role in the design, development, administration, scoring, use, and interpretation of the SAT Suite and the associated scores. It is also discussed more in-depth in a chapter on fairness, which provides the reader with a summary of fairness in one location (see Chapter 2: Fairness).

Reliability can be thought of as a prerequisite condition for the interpretation of test scores to be valid for a specific use. Reliability in this manual is further specified when discussed in the Reliability section, but it most generally refers to the " . . . consistency of the scores across instances of the testing procedure . . . " (AERA/APA/NCME, 2014, p. 33). In this manual, we outline the measures taken to ensure that the SAT and PSAT-related assessment scores are reliable across multiple forms of the assessment over time and in different contexts, and we document the reliability estimates of the assessment scores (see Section 6.4: Reliability).

When we refer to validity in this manual, it is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA/APA/NCME, 2014, p. 11). This means that if a test score has multiple uses and interpretations (e.g., admission to an institution *and* placement into coursework), each distinct use and interpretation must be validated. Note that validity is not a property of the test itself but rather refers to the interpretation of test scores for a specific use. The issue of validity is woven throughout this manual, as most processes and procedures discussed in this manual play a role in ensuring that the SAT and the PSAT-related assessments have scores that are valid for their intended uses. There is also a distinct chapter on validity that concludes this manual, where we outline the intended interpretations of specific uses of SAT Suite scores and provide the related validity evidence available for those uses (see Chapter 7: Validity).

1.2 Brief History of Development

The redesign of the SAT was announced on March 5, 2014, as part of the College Board Readiness and Success System, a system designed to make it easier for students to navigate a path through high school, college, and career. The system includes a suite of assessments with extensive actionable reporting, focused practice activities, and college and career information and opportunities for students. The redesigned assessments are available at multiple grade levels, all vertically aligned and scaled to provide educators and students with actionable feedback about students' college and career readiness from eighth grade through graduation. The College Board offers the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 as grade-appropriate assessment options for middle and high school students.

Background on Assessment

In 1926, 8,040 young men took what was then called the "Scholastic Aptitude Test" at its first administration. The 1926 version of the SAT bears little resemblance to the current test. It contained nine subtests: seven with verbal content and two with mathematical content. Beginning in 1930, the SAT was split into two sections, one portion designed to measure "verbal aptitude" and the other to measure "mathematical aptitude."

In 1959, the College Board created the PSAT (then called the Preliminary Scholastic Aptitude Test) to provide students with a less expensive test ($1 compared to the $7 SAT) to aid them in preparation for the SAT. Since that time, major strides have been made in fully developing the student-focused expansion of the role of the PSAT, which joined forces with National Merit Scholarship Corporation and became the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT) in 1971.

In fall 2008, the College Board field-tested the ReadiStep assessment, designed to measure the skills and knowledge necessary for eighth and ninth graders to be considered on track for college readiness. This assessment became the first step in the College Board's College and Career Readiness Pathway—which included the PSAT/NMSQT and the SAT. The design of ReadiStep's content was aligned with both the PSAT/NMSQT and the SAT. The PSAT 8/9, which replaced ReadiStep in the College Board Readiness and Success System, is administered in the eighth and/or ninth grades. For students, the PSAT 8/9 is the earliest opportunity they have to engage with the SAT Suite of Assessments.

In spring 2016, the PSAT 10 was launched as part of the SAT Suite of Assessments. The PSAT 10 is essentially the same test as the PSAT/NMSQT but is delivered in the spring rather than the fall of a given school year.

In 2009, the College Board began offering the SAT during the school day for eligible district and state partners. The SAT School Day program enables eligible states and districts to create a unique opportunity for all of their juniors or seniors to take the SAT in their home schools. SAT School Day provides encouragement for all students to pursue a college education and offers improved access and convenience to meet college admission testing requirements. There are also school day administrations for the PSAT/NMSQT, as well as a fall Saturday test date.

In keeping with best practices (AERA/APA/NCME, 2014), the SAT has been reconfigured several times. With each redesign, a variety of considerations were taken into account, including fairness issues, scaling issues, cost, public perception, face validity, changes in the test-taking population, changes in patterns of test preparation, changes in the curriculum and college expectations, and changes in the college admission process. Each redesign was intended to make the test more useful to students, teachers, high school counselors, and college admission staff.

In the most recent redesign of the SAT Suite, the College Board carefully examined what the best available evidence indicated were the essential prerequisites in reading, writing, language, and mathematics for readiness for and success in postsecondary education. This evidence, along with extensive feedback from our colleagues in K–12 and higher education, was critical to shaping the design of the assessment.

Reason for Changes

The College Board is committed to an opportunity agenda that is focused on propelling students into the opportunities they have earned in high school. One of the major components of this agenda is the redesign of the SAT Suite of Assessments. Drawing on extensive input and advice from our members, partner organizations, and postsecondary and K–12 experts, we determined that the new SAT Suite needed to address three challenges.

First, the tests must provide to higher education a more comprehensive and informative picture of student readiness for college-level work while sustaining, and ideally improving, the ability of the SAT Suite to predict college success. Second, the tests must be more clearly and transparently focused on the knowledge, skills, and understandings that the best available research indicates are essential for college and career readiness and success. Third, the tests must reflect, through their questions and tasks, the kinds of meaningful, engaging, and challenging work that students must undertake in the best high school courses being taught today, thereby creating a robust and durable bond between assessment and instruction. At the heart of these aims is the belief that all teachers and students must be empowered to focus on the real learning of vital knowledge, skills, and understandings through challenging, vibrant daily work, rather than being encouraged to cover vast swaths of material superficially or engage in narrow, short-term test preparation divorced from real learning. To these ends, the new SAT Suite has been designed for greater focus, relevance, and transparency while retaining the tradition of being a valuable predictor of college and career readiness and success.

Description of Key Features

In order to better reflect what the research reveals to be important, the College Board has updated the SAT and the PSAT-related assessments to focus on fewer topics that will be most relevant to postsecondary readiness and success (College Board, 2014). All components align to good classroom instruction, demanding deep thinking and rigorous analysis on questions grounded in real-world knowledge. The features of the SAT Suite of Assessments fall into eight categories, listed below. Several of these changes have been integrated into the tests to such an extent that they are key design elements of the revised assessments, as seen in Chapter 3: Test Development Procedures.

Eight Key Features

Words in Context. Instead of being asked to define obscure and seemingly random words of the kind commonly called "SAT words," test takers encounter relevant words and phrases that derive their meanings from the contexts in which they are used. Test takers engage in close reading and demonstrate the best work of the classroom. The skills tested are broadly useful in numerous subjects and careers.

Command of Evidence. Test takers analyze material from a variety of content areas (literature and literary nonfiction, science, history, and social studies) and on career-related topics. They use textual evidence to support their answers and apply an understanding of how authors make use of evidence.

Essay Analyzing a Source (SAT only). After completing the other three tests, test takers who opt to take the SAT Essay have 50 minutes to compose a clear and cogent analysis of a text in response to a prompt common to every administration of the SAT. Essays are scored

on reading comprehension, argument analysis, and writing skills. Similar to college writing assignments, this task promotes the practice of reading a wide variety of arguments and analyzing how authors do their work as writers.

Math That Matters Most. In keeping with the redesign's philosophy of deeper focus on fewer topics, the Math Test focuses on four areas that research shows are essential for college readiness: Heart of Algebra, Problem Solving and Data Analysis, Passport to Advanced Math (all tests[3]), and Additional Topics in Math (all but PSAT 8/9).

Problems Grounded in Real-World Contexts. Test takers engage with questions grounded in the real world and directly related to the work performed in college and career. Both the Reading Test and the Writing and Language Test include literature and literary nonfiction, but they also feature charts, graphs, and passages like those that students are likely to encounter in science, social science, and other majors and careers. The Math Test features multistep applications to solve problems in science, social science, career scenarios, and other real-life contexts.

Analysis in Science and in History/Social Studies. Across all components of the assessment, test takers are asked to apply their reading, writing, language, and math skills to answer questions in science, history, and social studies contexts. They will use these same skills throughout their lives to make sense of issues and topics.

U.S. Founding Documents and the Great Global Conversation. The U.S. founding documents, including the Declaration of Independence, the Bill of Rights, and the Federalist Papers, have helped inspire a conversation that continues to this day about the nature of civic life. Over time, authors, speakers, and thinkers from the United States and around the world, including Edmund Burke, Mary Wollstonecraft, and Mohandas Gandhi, have broadened and deepened the conversation around such vital matters as freedom, justice, and human dignity. Every time students take an assessment contained in the SAT Suite, they encounter a passage from one of the U.S. founding documents or a text from the Great Global Conversation. Our hope is to inspire a close reading of these rich, meaningful, often profound texts, not only as a way to develop valuable college and career readiness skills but also as an opportunity to reflect on and deeply engage with issues and concerns central to informed citizenship (College Board, 2015a). The use of these documents in the assessment is discussed further in Chapter 3: Test Development Procedures.

No Penalty for Guessing. The SAT Suite removes the correction for guessing that has been used to score the assessments in the past. Instead, test takers earn points for the questions they answer correctly. This move to rights-only scoring encourages test takers to give the best answer they have for every question, without fear of being penalized for making their best effort. Our general approach to scoring the SAT and the PSAT-related assessments is discussed further in Chapter 5: Interpretation and Application of Results.

Benefits

By reflecting the major shifts taking place in high school instruction, standards, and assessment, the assessments across the SAT Suite offer students, parents, admission officers, teachers, and counselors a better indicator of student progress toward becoming college and career ready. They also provide better information about students' strengths

[3] Passport to Advanced Math items appear in PSAT 8/9, but no subscore is given.

and weaknesses relating to the knowledge, skills, and understandings that are essential to college and career readiness and future success. In this manner, the SAT does a superior job of meeting its intended uses for its primary users.

For students and parents, the SAT Suite offers a more effective vehicle to showcase students' academic strengths and readiness for college and careers. Because it is more closely aligned to both high school instruction and post-high-school requirements, the SAT Suite serves as evidence of the hard work students have performed in high school, showing that challenging coursework and focused instruction can help provide opportunities for future success. Combined with high school grades and other factors that inform admission decisions, the SAT Suite gives students an opportunity to put their best foot forward in the admission process and demonstrate how well they have attained the knowledge, skills, and understandings necessary for postsecondary-level work. For more information on the relationship between challenging coursework and the SAT Suite, see later in this chapter.

For admission officers, the SAT provides a more detailed and comprehensive picture of each student's level of college readiness, helping colleges more easily identify students who are a good match for their institution and the programs of study it offers.

For K–12 educators and counselors, the SAT Suite offers clearer connections to classroom instruction, its questions and tasks more closely resembling the best of classroom teaching and better measuring the powerful knowledge, skills, and understandings needed in postsecondary education, work, and life. The assessments in the new SAT Suite offer an improved indicator of students' progress, through in-depth scores and reports that are designed to focus efforts on targeted areas of knowledge and skills with an integrated, personalized plan for practice and growth.

The assessments in the SAT Suite now contain the following benefits:

Rights-Only Scoring. As stated earlier in this chapter, scoring for the new assessments are based only on questions that are answered correctly. There is no point deduction for wrong answers, which encourages students to give the best answer for every question, rather than skip questions about which they are unsure.

Learning, Not Memorizing. The SAT Suite requires students to have a stronger command of fewer topics. Rote memorization and "cramming" to learn vocabulary that students soon forget are not a part of the SAT Suite. Students are asked to apply deep understanding of the skills and concepts most important for college and career readiness.

Connection to Classroom Learning and Experience. Students encounter assessments that are closely connected to their classroom experience and assessments that reward focused work and the development of valuable, durable knowledge, skills, and understandings. The questions and approaches students encounter are more familiar to them because they are modeled on the best work of classroom teachers.

Free Resources for Practice and Review. Students have access to free resources that introduce them to the SAT and the PSAT-related assessments and give them a chance to enhance their preparation with targeted review and authentic practice. The College Board has partnered with Khan Academy® to provide free practice materials that are personalized, interactive, and engaging to help students prepare for the SAT and the PSAT-related assessments. For more information about the College Board's partnership with Khan Academy, visit the SAT practice website, https://www.khanacademy.org/sat.

The benefits listed here represent those found in the PSAT-related assessments and the SAT. For a full list of the benefits offered by each test within the SAT Suite of Assessments, see Table A-1.1 in Appendix 1: Overview. These benefits represent the conceptual foundations for the SAT Suite. More information about the concordance between the old SAT and the new SAT can be found in Section 7.3.

The Role of High School Rigor on SAT Success

As stated earlier in this chapter, the academic rigor of a student's high school experience is an important component of success in college and career. Unfortunately, of 2016's high school graduates who took the old SAT exam while it was offered through January 2016 and who reported the courses they completed in high school, only roughly three out of four (78%) indicated that they had completed a core curriculum in high school, defined as at least four years of English and at least three years each of math, social science, and natural science. Far too many of our high school graduates aren't even taking the right *numbers* of courses in high school, much less the right *kinds* of courses. Moreover, of the students who completed a core curriculum, only 52% met the SAT benchmark, evidencing the need for more challenging core courses.

Far too many of our high school graduates are leaving K–12 education without the knowledge and skills they need to enter and succeed in some form of postsecondary education, which encompasses the majority of workforce training programs. Students spend far too much time in school focusing their daily work on texts that are not sufficiently complex, instead of focusing on evaluating evidence in reading and writing or on the command of math that matters the most to their future. They are not practicing the core work that matters the most in every course that they take. They battle barriers that few of them overcome in acquiring the skills needed to complete postsecondary education. As stated earlier in this manual, one of the foundations of the SAT is a greater connection to classroom learning and experience, so that students can acquire the knowledge that truly matters for college and career.

We must promote greater access to higher education by proactively working with middle schools and high schools to encourage all students to take the right number of core courses, to take the right kinds of courses, and to apply to four colleges.

Studies have demonstrated that students who take more challenging coursework in high school are more likely to be ready for college and career by the time they graduate from high school than students who take less challenging coursework. There is a strong relationship between challenging work in high school and SAT scores, as well as HSGPA, college enrollment, college freshman grade point average, and persistence to a second year of college (Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011).

We must promote greater access to challenging courses in high school by all students, particularly low-income, underrepresented students. Providing opportunities for access to challenging coursework in high school will increase equity in education and better prepare all students to succeed in college and career (Wyatt, Wiley, Camara, & Proestler, 2011).

In addition to offering and delivering challenging coursework, high schools play an important role in creating a college-going culture for students; students who attend high schools where there are high expectations, strong support for college attendance, and high participation in financial aid applications are more likely to apply, be accepted,

and enroll in four-year colleges that match their qualifications (Roderick, Coca, & Nagoka, 2011).

Improving college readiness can address the issue of inequity in education by increasing college graduation rates for all students, regardless of their ethnicity or household income levels. Students face inequitable access to academic rigor in high school, especially underrepresented minority students and low-income students (College Board, 2014). Providing access to rigor increases equity in education and better prepares students for success in college and career (Barry & Niu, 2013).

The Importance of Test Practice and SAT Success

In keeping with the College Board's previously stated belief in providing the same access to information and opportunities to all test takers, the College Board has partnered with Khan Academy to provide all test takers with free, personalized study resources for the SAT Suite of Assessments. This includes thousands of practice questions that are reviewed by the College Board, multiple practice tests written by the College Board, and personalized recommendations for instruction and practice to help students fill their knowledge gaps.

The College Board's partnership with Khan Academy will achieve several key goals. It is our hope that students who use Khan Academy will achieve improved scores on the SAT Suite of Assessments. More importantly, by providing free access to these Khan Academy materials we are giving *all* our students, including low-income and underrepresented students, the opportunity to achieve. Also, in keeping with the foundations of our redesign, the exercises available from Khan Academy provide students with productive practice aligned to classroom activities, matching the types of work they are currently completing in the classroom.

In order to continually monitor the usefulness of the test prep provided by Khan Academy, College Board will continue to use all available data and a range of research designs to evaluate and improve our practice resources.

1.3 Description of Content

As noted in the previous section, we based the SAT Suite of Assessments' test domain definitions on the highest-quality information and resources available about the essential requirements for college and career readiness and success. Scholarly research and empirical data derived from curriculum surveys conducted by the College Board and other organizations play an important role in informing these definitions. Our staff works with educational experts in examining the evidence and defining the domain of knowledge, skills, and understandings to be measured in accordance with each assessment's primary purpose and the claims associated with each assessment. We also prepare 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 test and question formats, in keeping with standards and best practices (AERA/APA/NCME, 2014). 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. A graphical summary of the scores corresponding to the test domains for each test in the SAT Suite will be presented later in this chapter.

Test Format

The SAT is organized into four tests: Reading Test, Writing and Language Test, Math Test, and an optional Essay, which is a direct-writing task. We designed each test 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 SAT's primary usage, assessing college and career readiness. The PSAT-related assessments contain the same basic content as those in the SAT, except the SAT is the only assessment that includes the optional Essay.

Each of the following sections is dedicated to a brief overview of each test format (Reading, Writing and Language, Math, and Essay), including the types of questions found on each test and a set of Quick Facts. Given the similarity in test content between the SAT and the PSAT-related assessments, the overview specifically covers the SAT tests but the description of test content can be applied to the PSAT-related assessments, although the specific content specifications and scales may be different. The overview of the Essay only applies to the SAT, as it does not appear on any of the PSAT-related assessments.

Tables 1.1, 1.2, and 1.3, provide an overview of the test format for SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9, respectively, while overview tables comparing content specifications across the integrated suite of assessments can be found in Appendix 1: Overview. For in-depth information about test specifications, see Chapter 3: Test Development Procedures.

SAT Reading Test

The SAT Reading Test is a 52-question, 65-minute assessment that is intended to collect evidence in support of a broad claim about student performance:

> Students can demonstrate college and career readiness proficiency in reading and comprehending a broad range of high-quality, appropriately challenging literary and informational texts in the content areas of U.S. and world literature, history/social studies, and science (College Board, 2014, p. 40).

The Reading Test is composed of a series of high-quality, previously published passages (including one pair of topically related passages) and associated multiple-choice questions. Passages represent the content areas of U.S. and world literature, history/social studies, and science, and exhibit a defined range of text complexity from early high school level to postsecondary entry. Questions ask test takers to derive information and ideas (stated or implied) from passages; to analyze passages rhetorically in terms of such features as point of view and purpose; and to draw connections between pairs of related passages and between informational graphics (tables, graphs, charts, and the like) and the passages they accompany.

Two areas of emphasis on the Reading Test—shared with the Writing and Language Test—are Command of Evidence and Words in Context. Test takers are, for example, routinely asked to cite the best evidence for the answer to a particular question or in support of a stated claim. Test takers are also routinely asked to determine the meaning of relevant words and phrases as they are used in particular contexts and to analyze how word choice influences the meaning and tone of passages.

Table 1.1: SAT Test Format

Test	Time Allotted (minutes)	Number of Questions/Tasks
Reading	65	52
Writing and Language	35	44
Math	80	58
Essay (optional)	50	1
Total	180	154
	(230 with Essay)*	(155 with Essay)

*These total times don't include breaks.

Table 1.2: PSAT/NMSQT and PSAT 10 Test Format

Test	Time Allotted (minutes)	Number of Questions/Tasks
Reading	60	47
Writing and Language	35	44
Math	70	48
Total	165*	139

*These total times don't include breaks.

Table 1.3: PSAT 8/9 Test Format (Fall and Spring)

Test	Time Allotted (minutes)	Number of Questions/Tasks
Reading	55	42
Writing and Language	30	40
Math	60	38
Total	145*	120

*These total times don't include breaks.

Quick Facts:

- All Reading Test questions are multiple choice and based on passages.
- Some passages are paired with other passages or accompanied by informational graphics, such as tables, graphs, and charts.
- Prior topic-specific knowledge is never tested.
- No mathematical computation is required.
- The Reading Test yields a test score on a scale ranging from 10 to 40. The test also contributes to two subscores, each on a scale ranging from 1 to 15: (1) Command of Evidence and (2) Words in Context.

For a comparison of the Content Specifications for the Reading Test across the SAT Suite of Assessments, see Table A-1.2 in Appendix 1: Overview.

SAT Writing and Language Test

The SAT Writing and Language Test is a 44-question, 35-minute assessment that is intended to collect evidence in support of a broad claim about student performance:

> Students can demonstrate college and career readiness proficiency in revising and editing a range of texts in a variety of content areas, both academic and career related, for expression of ideas and for conformity to the conventions of standard written English grammar, usage, and punctuation (College Board, 2014, p. 58).

A series of well-written and well-edited passages and their associated multiple-choice questions constitute the Writing and Language Test. Passages, which are developed expressly for the assessment, represent the content areas of history/social studies, the humanities, and science, as well as career-related subjects. The passages exhibit a defined range of text complexity from early high school level to postsecondary entry. Questions ask test takers to make effective, contextually based revision and editing decisions in multiparagraph passages in order to improve the passages' topic development, organization, and rhetorical use of language, as well as to correct errors in standard written English. Some passages are accompanied by informational graphics, such as tables, graphs, and charts, with associated questions asking test takers to use the data represented in the graphics (e.g., to add or revise support for particular points or to recognize and correct inaccuracies in a writer's interpretation of data).

Two areas of emphasis on the Writing and Language Test—shared with the Reading Test—are Command of Evidence and Words in Context. Test takers are routinely asked to establish or refine the central points of passages; add, revise, or delete supporting material; improve the focus of passages; and relate information presented quantitatively to information presented in the text. Test takers are also routinely asked to enhance the rhetorical effectiveness of language by making contextually based choices that improve precision and concision; maintain consistency in style and tone or achieve particular stylistic effects; and combine sentences and rearrange sentence elements in ways that improve clarity and accomplish specified rhetorical goals, such as placing emphasis on a main point rather than a subordinate point.

Quick Facts:

- All Writing and Language Test questions are multiple choice and based on passages.
- Some passages are accompanied by informational graphics, such as tables, graphs, and charts.
- Prior topic-specific knowledge is never tested.
- No mathematical computation is required.
- The Writing and Language Test yields a test score on a scale ranging from 10 to 40. The test also yields two subscores, each on a scale ranging from 1 to 15: (1) Expression of Ideas (topic development, organization, effective language use) and (2) Standard English Conventions (sentence structure, usage, punctuation). The test also contributes to two other subscores, each on a scale ranging from 1 to 15: (1) Command of Evidence and (2) Words in Context.

For a comparison of the content specifications for the Writing and Language Test across the SAT Suite of Assessments, see Table A-1.3 in Appendix 1: Overview.

SAT Math Test

The SAT Math Test is a 58-question, 80-minute test that is intended to collect evidence in support of the following claim about student performance:

> Students have fluency with, understanding of, and the ability to apply the mathematical concepts, skills, and practices that are most strongly prerequisite and central to their ability to progress through a range of college courses, career training, and career opportunities (College Board, 2014, p. 132).

The Math Test assesses four content areas: Heart of Algebra, Problem Solving and Data Analysis, Passport to Advanced Math, and Additional Topics in Math.[4] The test covers all mathematical practices, with an emphasis on problem solving, modeling, using appropriate tools strategically, and looking for and making use of structure to do algebra. Test takers are asked to demonstrate their command of the mathematics skills and knowledge that are most provably useful in a range of college courses and career environments. This is in keeping with the new assessment's commitment to a deeper focus on a few important topics. Along the same lines, this includes math addressing real-world situations and problems and multipart applications of this core of useful math.

The Math Test has two portions, one for which test takers are allowed to use calculators to solve the problems ("calculator portion") and the other for which test takers are not allowed to use calculators ("no calculator portion"). Most questions are multiple choice, but some require student-produced responses (SPR—"grid-in"). Test takers need to exhibit command of mathematical practices, fluency with mathematical procedures, and conceptual understanding of mathematical ideas. The problems on each Math Test explore the full dynamic range of each content area through precisely crafted questions that emphasize the use of math in unlocking insights and solving problems.

[4] Items under Additional Topics in Math contribute to the total Math score but don't contribute to a subscore within the Math Test.

Quick Facts:

- Most math questions are multiple choice, but some are student-produced responses (SPR—"grid-in").
- Calculators are allowed on one of two portions of the Math Test.
- Some parts of the test present test takers with a scenario and then ask several questions about it.
- The Math Test reports one overall test score ranging from 10 to 40 and three subscores: (1) Heart of Algebra, (2) Problem Solving and Data Analysis, and (3) Passport to Advanced Math, each on a scale ranging from 1 to 15.

For a comparison of the content specifications for the Math Test across the SAT Suite of Assessment, see Table A-1.4 in Appendix 1: Overview.

SAT Essay (Optional)

As with the other tests in the suite, the SAT Essay is intended to collect evidence in support of a broad claim about student performance:

> Students can demonstrate college and career readiness proficiency in producing a cogent and clear written analysis using evidence drawn from an appropriately challenging source text written for a broad audience (College Board, 2014, p. 69).

To that end, the optional 50-minute SAT Essay is designed to determine whether students can demonstrate college and career readiness proficiency in reading, analysis, and writing by comprehending a high-quality source text and producing a cogent and clear written analysis of that text supported by critical reasoning and evidence drawn from the source. The Essay is offered at the conclusion of the required tests (Reading, Writing and Language, and Math). Students taking the SAT during a national administration may choose not to take this portion of the assessment, and some postsecondary institutions may choose not to require it. Unlike much standardized direct-writing assessment, the Essay is not designed to elicit test takers' subjective opinions. Instead of asking them simply to emulate the form of evidence used by drawing on their own experiences or imaginations, the Essay requires test takers to make purposeful, substantive use of textual evidence in a way that can be objectively evaluated. In keeping with the test's emphasis on relevant knowledge that students will continue to encounter throughout their education, this task is designed to promote the practice of reading a wide variety of arguments and analyzing how authors do their work as writers.

The Essay prompt remains consistent for all administrations of the SAT; only the passages on which the test takers base their responses change. For the Essay, test takers are asked to explain how the author of the accompanying passage builds an argument to persuade an audience. Test takers are informed that they may analyze such aspects of the passage as the author's use of evidence, reasoning, and stylistic and persuasive elements but that they may also or instead choose other features to analyze; test takers are further advised that, in all cases, they should center their discussion on those aspects that are the most salient to the passage in question. Responses shouldn't focus on whether test takers agree or disagree with the claims made in the passage but rather on how the author builds an argument to persuade an audience. This approach more closely mirrors college writing assignments.

In broad terms, Essay responses are evaluated across three dimensions: demonstrated comprehension of the source text (Reading), the quality of analysis of that source text

(Analysis), and the quality of the writing in the response (Writing). For more information on how the optional Essay is scored, see Chapter 5: Interpretation and Application of Results.

Quick Facts:

- The Essay is optional.
- The Essay task remains consistent for all administrations of the SAT; only the source texts on which test takers base their responses change.
- The Essay requires test takers to make purposeful, substantive use of textual evidence in a way that can be evaluated objectively.
- Responses are assessed along three analytic scoring dimensions: Reading, Analysis, and Writing. Each dimension receives a score on a scale ranging from 2 to 8, the combination of two independent raters' scores on a 1–4 scale.
- The Essay scores aren't combined with each other or with the scores from any other portion of the test.

For the content specifications of the SAT Essay, see Table A-1.5 in Appendix 1: Overview.

Brief Overview on Scoring

Figures 1.1 through 1.3 provide an overview of how the tests in the SAT Suite are scored (more information on scoring is included in Chapter 5: Interpretation and Application of Results) and how the components of each assessment relate to one another. It should be noted that the scores on the PSAT-related assessments are on the same vertical scale as the SAT and differ only by the minimum and maximum scores reported on this vertical scale.

Figure 1.1: SAT scoring guide

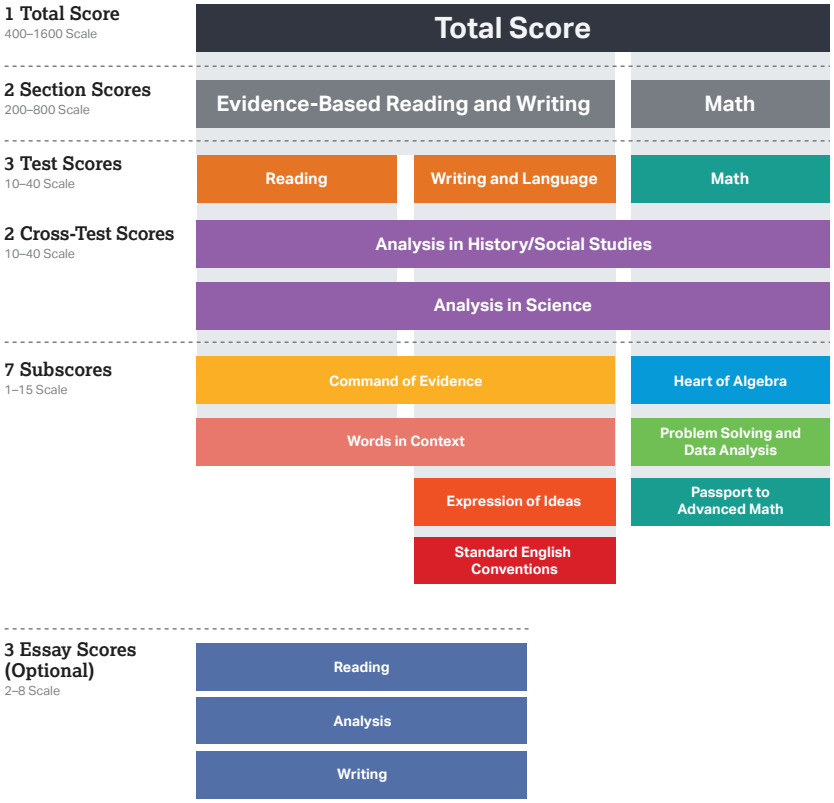


Figure 1.2: PSAT/NMSQT and PSAT 10 scoring guide

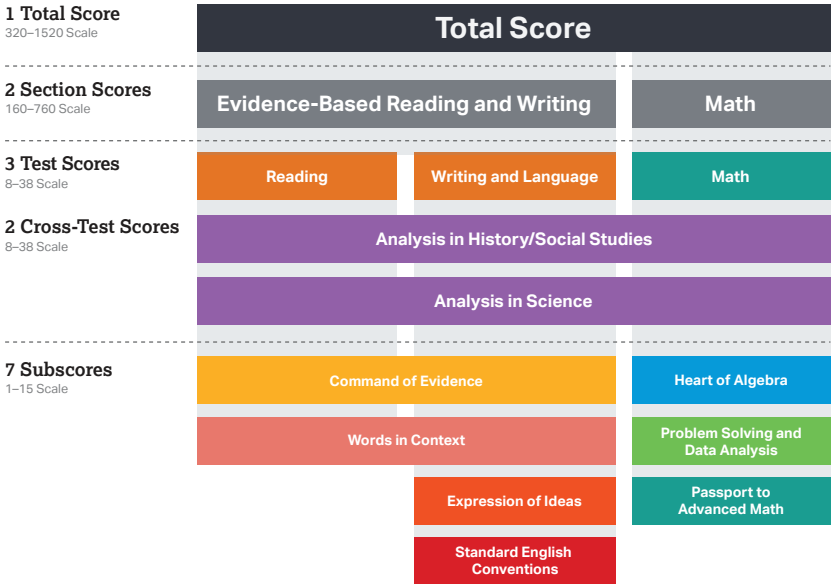
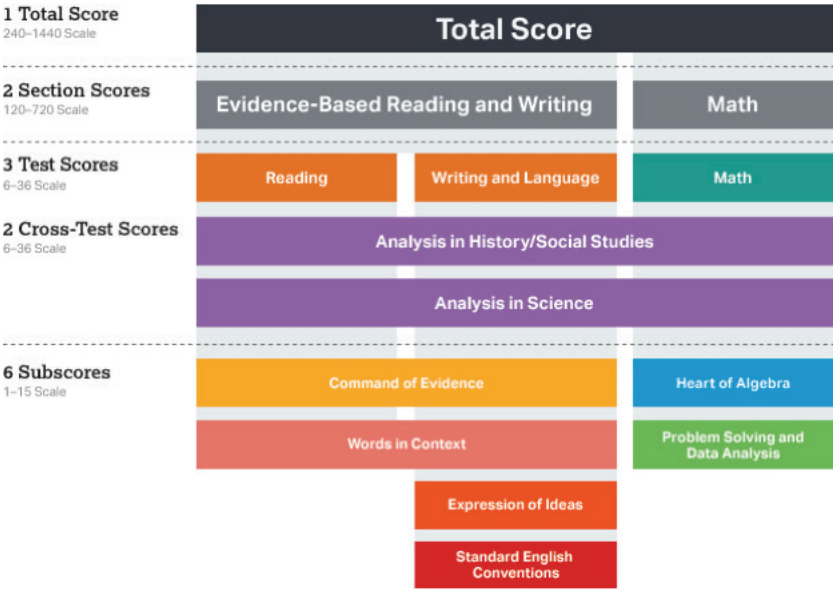


Figure 1.3: PSAT 8/9 scoring guide



Quick Facts:

SAT

- A Total Score, on a scale ranging from 400 to 1600, is the sum of the two section scores (1) Evidence-Based Reading and Writing and (2) Math.
- The SAT reports two section (domain) Scores: (1) Evidence-Based Reading and Writing, which is the sum of the Reading Test score and the Writing and Language Test score multiplied by 10, and (2) Math, which is the Math Test score multiplied by 20. Each of the two section scores is reported on a scale ranging from 200 to 800.
- The SAT reports three test scores in the range of 10–40; (1) Reading, (2) Writing and Language, and (3) Math
- The SAT reports two cross-test scores in the range of 10–40: (1) Analysis in History/Social Studies and (2) Analysis in Science, which are based on selected questions in the SAT Reading, Writing and Language, and Math Tests.
- The SAT reports seven subscores; two from the Reading Test and the Writing and Language Test, two from the Writing and Language Test, and three from the Math Test.
- The SAT report subscores in the range of 1–15 that offer feedback in the following skill areas:
  - Command of Evidence
  - Words in Context
  - Expression of Ideas
  - Standard English Conventions
  - Heart of Algebra
  - Problem Solving and Data Analysis
  - Passport to Advanced Math
- The optional Essay reports three scores in the range of 2–8; (1) Reading, (2) Analysis, and (3) Writing. The scores for the Essay are reported separately and aren't factored into the section scores.

PSAT/NMSQT and PSAT 10

- A Total Score, the scale ranges for the PSAT/NMSQT and PSAT 10 scores are 320–1520 for the total score.
- The PSAT/NMSQT and PSAT 10 reports two section Scores: (1) Evidence-Based Reading and Writing, which is the sum of the Reading Test score and the Writing and Language Test score, and (2) Math, which is the Math Test scores (including both the Math Test – Calculator and Math Test – No Calculator portions). Each of the two section scores is reported on a scale ranging from 160 to 760.
- The PSAT/NMSQT and PSAT 10 reports three test scores in the range of 8–38; (1) Reading, (2) Writing and Language, and (3) Math

- The PSAT/NMSQT and the PSAT 10 report two cross-test scores in the range of 8–38. These scores represent student performance on items across the three tests that were in the domains of either: Analysis in History/Social Studies and Analysis in Science.
- The PSAT/NMSQT and PSAT 10 report subscores in the range of 1–15 that offer feedback in the following skill areas:
  - Command of Evidence
  - Words in Context
  - Expression of Ideas
  - Standard English Conventions
  - Heart of Algebra
  - Problem Solving and Data Analysis
  - Passport to Advanced Math

PSAT 8/9

- A Total Score, the scale ranges for the PSAT 8/9 scores are 240–1440 for the total score.
- The PSAT 8/9 reports two section Scores: (1) Evidence-Based Reading and Writing, which is the sum of the Reading Test score and the Writing and Language Test score, and (2) Math, which is the Math Test scores (including both the Math Test – Calculator and Math Test – No Calculator portions). Each of the two section scores is reported on a scale ranging from 120 to 720.
- The PSAT 8/9 reports three test scores in the range of 6–36: (1) Reading, (2) Writing and Language, and (3) Math.
- The PSAT 8/9 reports two cross-test scores in the range of 6–36. These scores represent student performance on items across the three tests that were in the domains of either History/Social Studies or Science: Analysis in History/Social Studies and Analysis in Science.
- The PSAT 8/9 report subscores in the range of 1–15 that offer feedback in the following skill areas:
  - Command of Evidence
  - Words in Context
  - Expression of Ideas
  - Standard English Conventions
  - Heart of Algebra
  - Problem Solving and Data Analysis

CHAPTER 2

Fairness

Fairness is a central tenet of all College Board assessments, one with a number of wide-ranging implications. To this end, fairness is a recurrent theme across this entire manual and isn't confined to this one chapter. Section 2.1 provides an overview of what fairness is as it applies to our assessments and the places in this manual that the various aspects of fairness are discussed at length. This chapter then provides an in-depth look at a few of the important ways our assessments demonstrate fairness. This includes a consideration of the items and forms that make up our assessments (Section 2.2) and the manner in which students with disabilities receive accommodations during a test administration (Section 2.3). Upcoming versions of this technical manual will touch on the subgroup differences in PSAT-related assessment scores and SAT scores (Section 2.4) and the differential validity and prediction analyses for SAT scores with first-year grade point average (FYGPA) (Section 2.5).

2.1 What Is Fairness?

The College Board believes in providing all of our students with a fair opportunity to demonstrate their standing on our assessments. Fairness can be seen as both the equitable treatment of all test takers in the test administration and the equal measurement quality across subgroups and populations. Best practices and 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 equally for all intended test takers and test-taking populations. It should be administered in a manner that is fair and equitable for all test takers, regardless of gender, race/ethnicity, and other special considerations. In order to accomplish this goal, several aspects of fairness should be addressed when developing and administering an assessment.

One aspect of fairness is item fairness. Test takers demonstrating the same level of achievement in a content area should have similar chances of answering each question correctly, regardless of gender or of race/ethnicity. One major approach to ensuring that the items and the assessment are fair is to investigate differential item functioning, or DIF, wherein two groups are compared on individual item performance to determine if one group is systematically performing differently than the other group. These calculations, in conjunction with expert panel reviews, are used to ensure that individual items aren't performing differently for different subgroups and whether some element *unrelated* to those constructs being measured is affecting the scores. Section 2.2 provides more information about this process.

Fairness extends beyond item performance and test construction and is strongly tied to validity. Standards 3.6–3.8 (AERA/APA/NCME, 2014) address the notion that fair assessments ensure validity of test score interpretations across all groups within the test taker population. For instance, it is the responsibility of the test developer to investigate and ensure that test scores maintain the same meaning across various subgroups. In this case,

test scores shouldn't provide different criterion predictions for different subgroups. This evidence is obtained by investigating predictive validity by subgroup; often called differential validity and prediction (see Section 2.5 or Chapter 7).

Fairness also involves equality in test administration across all groups of test takers. For instance, detailed procedures are specified by the College Board to ensure that the assessment is administered uniformly across all testing sites in a fair and equitable manner. Security measures are also set in place to ensure that no test taker or group of test takers obtain access to information or opportunities that allow them to attain scores by fraudulent means and jeopardize the validity of the results of the assessment (AERA/APA/NCME, 2014). More information on these procedures is in Sections 4.1 and 4.2.

In keeping with the importance of providing the same access to information and opportunities to all test takers, the College Board has partnered with Khan Academy to provide free, personalized study resources for the SAT Suite of Assessments to all test takers. This includes thousands of practice questions that are reviewed and approved by the College Board, multiple practice tests written by the College Board, and personalized recommendations for instruction and practice to help students fill their knowledge gaps. To ensure that as many students as possible can take advantage of these free online resources, the College Board is also collaborating with community-based organizations, including the Boys & Girls Clubs of America. For more information on the study materials and information provided, see Section 1.2.

Some test takers may need additional support in order to complete the assessment and/or obtain valid test scores. Testing programs may address this concern by offering test takers accommodations for testing, such as large print, braille, or extra time. Standards 3.9 through 3.14 discuss the responsibility of test developers to develop and provide test accommodations and the appropriate use of said accommodations (AERA/APA/NCME, 2014). These accommodations must be documented and also must allow for testing to be conducted without changing the construct or constructs being measured, such that scores maintain their meaning across all subgroups, as well as accommodated and non-accommodated students. For more information on accommodations for the SAT Suite of Assessments, see Sections 2.3 and 3.5.

2.2 Fairness Reviews of Items, Forms, and Prompts for the SAT and the PSAT-Related Assessments

The College Board is committed to providing our students with fair assessments. This fairness is attended to at every stage of the test development process, from test design to delivery of score reports. To this end, we take great care to measure only the skills and knowledge defined in our test specifications. Letting other factors influence the questions and test forms could cause the performances of test takers to be affected in ways that are unrelated to the constructs intended to be measured and, consequently, result in scores that wouldn't provide an accurate measure of student achievement.

Fairness Reviews Prior to Pretesting

Prior to pretesting, all questions are reviewed by external, independent reviewers who are asked to evaluate each question according to a set of criteria for content accuracy and fairness. These reviewers are typically active classroom teachers drawn from across the nation at both the secondary and postsecondary levels and are deeply familiar with both

the student population of interest and the nature and purpose of the assessment. Fairness reviewers are charged with helping ensure that test questions and stimuli are broadly accessible to the wide-ranging student population that takes the SAT Suite of Assessments, that the questions are clearly stated and unambiguous in their intent, and that the questions don't offer unfair advantages to some test takers. The guidelines provided to our fairness reviewers as they review test questions and stimuli are summarized in this chapter.

Topics to Avoid. Topics that are highly sensitive or controversial among various population groups, generally not required by the constructs being measured by the tests, should typically be avoided.

Portrayal. All population groups should be portrayed fairly, authentically, and with respect. No group should be depicted in a demeaning (including self-demeaning) way by test materials.

Stereotyping. Instances of stereotyping should be avoided, whether the stereotype is negative or "positive." Examples include portraying all women as homemakers and cooks; girls as ballerinas and boys as athletes; African Americans as people who live in the inner city or who excel exclusively in sports; Asian Americans as particularly gifted in math; Native Americans as particularly attuned to nature; or older people as feeble, slow, incompetent, or dependent. All groups should be represented as having the broadest possible range of points of view, talents, interests, and aspirations.

Group Identification. For the appropriate terminology for given population groups, it is generally best practice to use the terminology that respective group members themselves prefer. Outmoded group identifications are sometimes acceptable in certain historical or cultural contexts (e.g., *The Journal of Negro History*, National Association for the Advancement of Colored People). Men and women should be identified in comparable ways (e.g., both by first name or both by last name).

Language. Given that different groups of test takers have different rates of exposure to them, foreign words and phrases, slang and dialect, and idiomatic expressions should be avoided in testing materials unless their use is pertinent to the construct.

Ethnocentrism. Test materials shouldn't treat aspects of U.S. or Western political systems, society, cultures, or values as universal when they are specific to those regions.

Regionalism. Regional differences are often subtle. Referring to "yard sales" or "garage sales," for example, could be confusing for urban test takers. (Similarly, "stoop sales" could be confusing to rural or suburban test takers.) Generic terms are, therefore, generally preferable to region-specific ones. For example, a term such as "soft drink" is usually better than "pop," "soda," or "coke." On the other hand, reviewers shouldn't make overly restrictive assumptions about test takers' experiences or ability to make inferences from text. Even test takers who have never seen snow, lightning bugs, or a subway, for example, are likely to understand these concepts if the test materials explain them well enough.

Testing Context. The testing situation can be stressful for test takers. Therefore, special care should be taken to avoid content and contexts that seem likely to trigger or add to that stress. Highly controversial topics (such as abortion or the death penalty) that might potentially be discussed safely in a classroom environment, in which a teacher can mediate and students have the time and space to express their ideas and feelings, are often inappropriate for a testing context, during which students are timed and have no emotional outlet.

Pretesting Questions

All questions are pretested on a motivated sample of test takers that resembles the population of interest and is sufficient in size to allow the 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 ensure that test takers from different racial/ethnic and gender groups don't differentially respond to the questions. The questions are administered to test takers in test administrations like those in which the SAT Suite is given. The data from 1,000 to 3,000 test takers responding to each question are used to evaluate the performance of the questions. Once questions and tasks have been pretested and the statistics associated with them have been computed, the materials are reviewed by measurement and content specialists (including active classroom teachers at both the secondary and postsecondary levels) for content accuracy, fairness, statistical discrimination, difficulty, and differential performance among groups of tested students.

DIF Analyses

Analyses of DIF are conducted on the items at the pretest stage to identify questions that may function differently for members of different groups. 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 gender or race/ethnicity. DIF occurs when individuals with similar scores on an assessment differ notably in their performance on a specific test question by student subgroup. The presence of DIF indicates that a question functions differently for individuals from one subgroup from the way it functions for those of another subgroup who are at the same individual achievement level. Items exhibiting high levels of DIF may be measuring factors irrelevant to an assessment (such as culture) or more than one dimension for which the two groups have different strengths.

DIF analyses begin by examining any differences in the performance on each individual question of two comparable achievement level groups, referred to as the reference group and the focal group. Questions having extreme values of DIF, or those questions appearing to favor one group over another for test takers of the same level of achievement, undergo further review to determine whether some aspect of what the question is measuring is particularly related to subgroup membership and irrelevant to the dimension being measured. When a question is identified as exhibiting such characteristics, it is either revised and re-pretested or eliminated. For more information on DIF as it relates to the SAT Suite, see Chapter 3: Test Development Procedures.

Operational Forms Content and Fairness Reviews

Once test forms are initially constructed, the forms undergo multiple internal and external content and fairness reviews prior to finalization and preparation for publication. External review committee members are typically active classroom teachers drawn from across the nation and from both the secondary and postsecondary levels.

2.3 Test Accommodations to Remove Construct-Irrelevant Barriers

An important way that the College Board seeks to provide a fair testing environment for all test takers is allowing students with disabilities to take the tests in the SAT Suite with the accommodations they need (College Board, 2015d). This practice ensures that, when appropriate and possible, we are removing construct-irrelevant barriers that can interfere with a test taker accurately demonstrating their true standing on a construct (AERA/APA/NCME, 2014). A construct-irrelevant barrier is any factor(s) 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 particular uses.

The accommodations offered by the College Board serve to remove unfair disadvantages for those students with disabilities who have been approved to use accommodations on College Board assessments. In keeping with 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 forms 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. For a list of some of the available College Board testing accommodations, see Section 3.5. Accommodations are not limited to those listed, as the College Board considers any accommodation for any documented disability, as long as a student qualifies for testing accommodations.

Although numerous accommodations are possible, students with disabilities don't qualify automatically for testing accommodations, but must submit a request for approval by the College Board. Beginning Jan. 1, 2017, the vast majority of students who are approved for and using testing accommodations at their school through a current Individualized Education Program (IEP) or 504 Plan have those same accommodations automatically approved for taking College Board assessments. Most private school students with a current, formal school-based plan that meets College Board criteria also have their current accommodations automatically approved for College Board assessments.

In those instances where a student doesn't qualify for automatic approval through the school verification process, the request and documentation are reviewed by the College Board's Services for Students with Disabilities (SSD). In general, students approved by SSD to receive College Board testing accommodations meet the following criteria:

Student has a documented disability. Examples of disabilities include, but aren't limited to, visual impairments; learning disorders; physical and medical impairments; and motor impairments. Students 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 a College Board assessment is impacted. The disability must result in a relevant functional limitation that impacts the student's ability to participate in College Board assessments. For example, students whose disabilities result in functional limitations in reading, writing, and sitting for extended periods may need accommodations on College Board assessments, given the components of many of our tests and the manner in which assessments are generally administered.

Requested accommodation is needed. The student must demonstrate the need for the specific accommodation requested. For example, students requesting extended time should have documentation showing that they have difficulty performing timed tasks, such as testing under timed conditions.

Approved accommodations remain in effect until one year after high school graduation (with some limited exceptions) and can be used on the SAT, SAT Subject Tests, PSAT/NMSQT, PSAT 10, and the AP Exams. Students **do not need** to request accommodations from the College Board for subsequent assessments taken during this eligibility period. Approval from the College Board isn't required for the use of accommodations on the PSAT 8/9 assessment at this time. The accommodations available for other assessments may also be used on the PSAT 8/9. More information about the availability of accommodations and the procedures for requesting them prior to testing can be found at College Board's SSD website, https://www.collegeboard.org/students-with-disabilities.

2.4 Subgroup Differences on the SAT and the PSAT-Related Assessments

When reporting on student performance across the SAT Suite of Assessments, the College Board will include information on the performance of important demographic subgroups. Results of subgroup performance analyses will be shared when cohort data are available.

2.5 Differential Validity and Prediction Analyses for the SAT with FYGPA

The College Board will launch an SAT differential validity and prediction study, examining students in the entering college class of fall 2017, the first full cohort to be admitted with the new SAT. These students will complete one year of college and in that following year, we will be able to study the relationship between SAT scores and first-year college performance by student subgroups such as sex, race/ethnicity, socioeconomic status, and best language. When the results of this study are available, they will be shared online.

CHAPTER 3

Test Development Procedures

The College Board seeks to make the SAT Suite of Assessments deeply reflect the work that students need to do to be ready for and successful in college and their career paths. The individual questions and tasks and the tests as a whole reinforce enriching and valuable schoolwork, reflect a deep commitment to craft, and can be used by states and teachers to help define the level of rigor required for students to be college and career ready by no later than the end of high school.

The College Board works with various committees throughout the test design and development process to ensure the highest-quality assessments, ones that serve students well as they work to become college and career ready. The College Board's academic advisory committees and test development committees, which include secondary and postsecondary classroom teachers, advise throughout the development process, help define what constitutes academic preparation needed for college, design the test, help to develop item specifications, and review every question multiple times before it's placed on an operational test form. When reviewing test questions and forms, the test development committees help to ensure that the questions are measuring important and nontrivial skills, that the skills the questions are assessing align well with the test specifications in terms of content and rigor, that all test questions are fair to all students, and that the questions are written in a way that models good instruction for the teacher and productive practice for the student.

The College Board works with K–12 teachers and postsecondary instructors who teach entry-level courses to develop challenging items for the test that represent the kinds of tasks that students need to be able to perform if they are to be successful in higher education. Using items written and/or reviewed by K–12 and postsecondary instructors helps to ensure that the items are relevant to the work students do in challenging classrooms and reflect the kinds of tasks that students will be expected to perform in college.

Guiding Principles of the College Board's Test Development Process

To achieve the vision outlined above, every test form for the SAT Suite of Assessments is developed with care and expertise at every stage of the process. To that end, the College Board has implemented a test development process that helps ensure that our SAT Suite of Assessments questions and tasks are:

- 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 knowledge, skills, and understandings 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 with the help of classroom teachers 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 and again prior to operational form administration.
- 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.

The test development process for the SAT Suite of Assessments is based on these guiding principles. This chapter details the steps in that process, from the establishment of test specifications until the test is ready to be administered to test takers. Section 3.1 discusses the Test Specifications for each test in the SAT Suite. Section 3.2 details the creation, review, and analysis of the items that make up the tests. Section 3.3 addresses how those items are then assembled into test forms. Section 3.4 covers the test development procedures as they apply to the optional SAT Essay. Section 3.5 details a number of the comparable forms of the tests that have been developed for test takers requiring certain accommodations. Finally, Section 3.6 discusses the certification and manufacturing of the test materials, prior to their distribution to students.

3.1 Test Specifications

Content Specifications for SAT Suite of Assessments Reading Tests

Overall Claim for the Tests

The SAT Suite Reading Tests are designed to collect evidence in support of a broad claim about student performance:

> Students can demonstrate college and career readiness proficiency, as defined by their grade level, in reading and comprehending a broad range of high-quality, appropriately challenging literary and informational texts in the content areas of U.S. and world literature, history/social studies, and science (College Board, 2014, p. 40).

Test Description

As stated previously, the basic aim of the SAT Suite Reading Tests is to determine whether students can demonstrate college and career readiness proficiency as defined by their grade level in comprehending a broad range of high-quality, appropriately challenging literary and informational texts in the content areas of U.S. and world literature, history/social studies, and science. The assessments comprise a series of passages and the multiple-choice (MC) questions associated with them, as shown in Table 3.1. To answer the questions, test takers must refer to what the passages say explicitly and use careful reasoning to draw supportable inferences from the passages. In some cases, topically related passages in history/social studies and in science are paired and accompanied by questions assessing whether test takers can draw important connections between the passages as well as comprehend each passage individually. In other cases, history/social studies and science passages are accompanied by one or more relevant graphical representations of data—tables, graphs, charts, and the like—and certain questions require students to interpret the graphic(s) and/or to synthesize information and ideas presented graphically with those in the associated passage. (Mathematical computation is, however, not required to answer these questions.)

Table 3.1: Number of Questions/Time Limits for the SAT Suite Reading Tests

SAT Suite Reading Tests	Time Limit	Number of Questions
SAT	65 minutes	52 questions
PSAT/NMSQT and PSAT10	60 minutes	47 questions
PSAT 8/9	55 minutes	42 questions

All passages are taken from high-quality, previously published sources; all graphics are either also taken from such sources or created for the test based on authentic, accurate data. Each prose passage is intended to represent some of the best writing and thinking in the field it represents. Literature selections come from classic and contemporary works by authors working in the United States and around the world. History/social studies selections include portions of U.S.-based founding documents and texts in the Great Global Conversation—engaging, often historically and culturally important works grappling with issues at the heart of civic and political life—and explorations of topics in the social sciences, including anthropology, communication studies, economics, education, human geography, law, linguistics, political science, psychology, and sociology (and their subfields). Science selections examine both foundational concepts and recent developments in the natural sciences, including Earth science, biology, chemistry, and physics (and their subfields).

The questions associated with the passages assess whether students understand information and ideas in these readings; are able to analyze texts rhetorically; and can synthesize across topically related passages as well as a passage and its accompanying graphic(s). Questions address substantive information and ideas in passages and graphics, and they are meant to reflect the kinds of questions one would encounter in a lively, rigorous, evidence-based discussion of the texts. The order in which questions are presented is also as natural as possible, with general questions about central ideas, themes, point of view, overall text structure, and the like coming early in the sequence (so that students can first build and demonstrate an understanding of the passage as a whole), followed by more localized questions about details, words in context, evidence, and the like. Answers are derived from what is stated or implied in the passages and graphics rather than on prior knowledge of the topics.

Numerous questions also address whether students are able to interpret the meaning of words and phrases in context and/or analyze how word choice influences meaning, shapes mood and tone, reflects point of view, or lends precision or interest. In addition, certain questions require analysis and synthesis of information and ideas presented in multiple, related passages or in a passage and its associated graphic(s). Across each test, test takers are asked to analyze passages and graphics in ways consistent with how texts are read in the content areas they represent, so that questions about a science passage, for example, might focus on hypotheses, experimentation, and data, while questions about a literature passage might focus on theme, mood, and characterization (although, again, topic-specific prior knowledge isn't assessed).

In conclusion, the SAT Suite of Assessments Reading Tests are challenging, carefully constructed assessments of comprehension and reasoning skill with an unmistakable focus on close reading of appropriately challenging passages and graphics in a wide array of subject areas.

Test Summary

For the full content specifications of the SAT Suite Reading Tests and a list of the dimensions of the SAT Suite Reading domain, see Table A-3.1 and Table A-3.2, respectively, in Appendix 3: Test Development Procedures.

Key Features of SAT Suite of Assessments Reading Tests

Four distinctive features of the SAT Suite Reading Tests are described below (the first two comprise subscores based on items from both the Reading Test and the Writing and Language Test in each of the SAT Suite of Assessments; for more information on subscores, see Chapter 1: Overview and Section 5.1):

- Emphasis on words in context
- Emphasis on command of evidence
- Inclusion of informational graphics
- Specified range of text complexity

For the Evidentiary Foundations of these features, see Section 7.2.

Words in Context. Reading Tests in the SAT Suite measure test takers' understanding of the meaning and use of words and phrases in the context of extended prose passages. These words and phrases are neither highly obscure nor specific to any one domain. They are words and phrases whose specific meaning and rhetorical purpose are derived in large part through the context in which they are used.

Command of Evidence. Reading Tests in the SAT Suite require test takers not only to derive information and ideas from a text but also in some cases to identify the portion of the text that serves as the best evidence for the conclusions they reach. In this way, test takers both interpret text and back up their interpretation by citing the most relevant textual support.

Informational Graphics. Reading Tests in the SAT Suite each include two passages that include one or two graphics (e.g., tables, graphs, and charts) that convey information related to the passage content. Test takers are asked to interpret the information conveyed in one or more graphics and/or to integrate that information with information in the text.

Text Complexity. Reading Tests in the SAT Suite include passages that span a range of text complexity levels from grades 6–8 to postsecondary entry. Test development staff make use of quantitative and qualitative measures of text complexity, feedback from subject-matter experts at the K–12 and postsecondary levels, and student performance data to make and refine decisions about the placement of passages within bands. These steps help ensure that the range of text difficulties represented on test forms are comparable.

Content Specifications for SAT Suite of Assessments Writing and Language Tests

Overall Claim for the Test

Like the other tests in the SAT Suite, the Writing and Language Tests are designed to collect evidence in support of a broad claim about student performance:

> Students can demonstrate college and career readiness proficiency, as defined by their grade level, in revising and editing a range of texts in a variety of content areas, both academic and career related, for expression of ideas and for conformity to the conventions of standard written English grammar, usage, and punctuation (College Board, 2014, p. 58).

Test Description

The Writing and Language assessments comprise a series of passages and the multiple-choice questions associated with them, as shown in Table 3.2 below. Some passages and/or questions are accompanied by one or more graphical representations of data—tables, charts, graphs, and the like—and certain questions require students to make revising and editing decisions to passages in light of information and ideas conveyed graphically. Mathematical computation is, however, not required to answer these questions.

All passages are written specifically for the tests so that errors (a collective term for various content-related, rhetorical, or mechanical problems) can be introduced that students must recognize and correct. The most common question format requires test takers to choose the best of three alternatives to an indicated part of the passage (often an underlined portion) or to determine that the version presented in the passage is the best option; other formats, however, are also used. All graphics are either taken from high-quality previously published sources or created for the test based on authentic, accurate data. In their base, "correct" form, passages are well-written essayistic prose pieces on topics in careers, history/social studies, the humanities, and science, with the core writing modes of argument, informative/explanatory text, and nonfiction narrative represented. Careers passages typically deal with trends, issues, and debates in major fields of work, such as information technology or healthcare. History/social studies passages discuss historical topics or topics in the social sciences, including anthropology, communication studies, economics, education, human geography, law, linguistics, political science, psychology, and sociology (and their subfields). Humanities passages delve into subjects in the arts and letters. Science passages explore concepts, research, and discoveries in the natural sciences, including Earth science, biology, chemistry, and physics (and their subfields).

Table 3.2: Number of Questions/Time Limits for the SAT Suite Writing and Language Tests

SAT Suite Writing and Language Tests	Time Limit	Number of Questions
SAT	35 minutes	44 questions
PSAT/NMSQT and PSAT 10	35 minutes	44 questions
PSAT 8/9	30 minutes	40 questions

The questions associated with the passages place students in the role of someone revising and editing the work of an unspecified writer. Test takers are, by turns, asked to improve the development, organization, and use of language in the passages and to ensure that the passages conform to conventions of standard written English grammar, usage, and punctuation. When passages and/or questions are accompanied by graphics, test takers are asked to draw connections between text and graphics—for example, to correct a passage's inaccurate interpretation of data presented in a table. Answers to all questions are anchored in the context of the passages. Neither rote recall of language rules nor context-free applications of grammar, usage, and mechanics conventions are tested; moreover, low-level recognition and labeling of errors is downplayed in favor of asking test takers to make authentic, context-based revising and editing decisions.

National curriculum surveys conducted by the College Board and others demonstrate that postsecondary instructors rate high in importance a student's ability to analyze an entire text and evaluate it in context (see Section 7.2 for more information). In accordance with this, the SAT Suite tests require test takers to answer questions based on extended-prose contexts rather than in isolation or in limited (e.g., single-sentence) contexts. Although some questions are answerable by referring to a single phrase, clause, or sentence, many others leverage the extended context the test's format makes available and require students to have an understanding of multiple sentences, one or more paragraphs, or the passage as a whole.

The range of rhetorical and conventions issues assessed on the Writing and Language Tests in the SAT Suite have been carefully delineated by the tests' blueprints to ensure that the matters deemed most relevant to future postsecondary success are emphasized in test questions (see Table A-3.3 in Appendix 3: Test Development Procedures). Furthermore, the Writing and Language Tests support the SAT Suite's focus on Command of Evidence and on Words in Context by allocating numerous questions to assess whether students can develop ideas effectively (e.g., by adding relevant supporting details or by maintaining or improving focus and cohesion) and use words carefully and with purpose (e.g., to improve precision or concision).

The Writing and Language Tests also exemplify the SAT Suite's emphasis on literacy across the curriculum by its inclusion of appropriately challenging passages in numerous content areas, including history/social studies and science. Moreover, the Writing and Language Tests reinforce the commitment of the SAT Suite to assessing quantitative literacy by including graphics and graphics-based questions. Much like the SAT Suite Reading Tests, the SAT Suite Writing and Language Tests present students with rigorous, carefully designed assessments of key literacy competencies needed for college and careers.

Test Summary

For the full content specifications of the SAT Suite of Assessments Writing and Language Tests and a list of the dimensions of the SAT Suite Writing and Language domain, see Table A-3.3 and Table A-3.4, respectively, in Appendix 3: Test Development Procedures.

Key Features of SAT Suite Writing and Language Tests

Three distinctive features of the SAT Suite Writing and Language Tests (shared with the SAT Suite Reading Tests) are described in the following text (the first two comprise subscores

based on items from both the Reading Test and the Writing and Language Test of each assessment; for more information on subscores, see Chapter 1: Overview and Section 5.1):

- Emphasis on Words in Context
- Emphasis on Command of Evidence
- Inclusion of informational graphics

For the evidentiary foundations of these features, see Section 7.2.

Words in Context. Writing and Language Tests in the SAT Suite measure the ability of test takers to apply knowledge of words, phrases, and language in general in the context of extended prose passages.

Command of Evidence. Writing and Language Tests in the SAT Suite measure test takers' capacity to revise a text to improve its development of information and ideas. To answer these questions, test takers must have a solid grasp of the content of the passage in question (although it's important to note that prior knowledge of the topic isn't expected).

Informational Graphics. Writing and Language Tests in the SAT Suite include one or more passages and/or questions that include one or more graphics (e.g., tables, graphs, or charts) that convey information related to the passage content. Test takers are asked to consider the information conveyed in these graphics as they make decisions about how and whether to revise a passage.

Content Specifications for SAT Suite of Assessments Math Tests

This following text describes the content, format, and distinctive features of the SAT Suite Math Tests as well as the skills they measure.

Overall Claim for the Test

The SAT Suite Math Tests are designed to collect evidence in support of the following claim about student performance:

> Students, as defined by their grade level, have fluency with, understanding of, and the ability to apply the mathematical concepts, skills, and practices that are most strongly prerequisite and central to their ability to progress through a range of college courses, career training, and career opportunities (College Board, 2014, p. 131).

Test Description

In keeping with the evidence about essential requirements for college and career readiness, the Math Tests in the SAT Suite of Assessments require a strong command of fewer, more important topics. Table 3.3 shows the number of questions and the time limits by each Math Test within the SAT Suite. To succeed on the Math Tests, test takers need to exhibit mathematical practices, such as problem solving and using appropriate tools strategically. The Math Tests in the SAT Suite also provide opportunities for richer applied problems.

The SAT Suite Math Tests have four content areas (the first two or three, depending on the assessment, are subscores; for more information on subscores, see Chapter 1: Overview and Section 5.1):

- Heart of Algebra
- Problem Solving and Data Analysis

- Passport to Advanced Math
- Additional Topics in Math

For the evidentiary foundations of these content areas, see Section 7.2.

Questions in each content area span the full range of difficulty and address relevant practices, fluency, and conceptual understanding.

Table 3.3: Number of Questions/Time Limits for the SAT Suite Math Tests

SAT Suite Math Tests	Time Limit	Number of Questions
SAT	80 minutes	58 questions
PSAT/NMSQT and PSAT 10	70 minutes	48 questions
PSAT 8/9	60 minutes	38 questions

Test Summary

For the full content specifications of the SAT Suite Math Tests and a list of the dimensions of the SAT Suite Math domain, see Table A-3.5 and Table A-3.6, respectively, in Appendix 3: Test Development Procedures.

The Math Tests cover all mathematical practices, with an emphasis on problem solving, modeling, using appropriate tools strategically, and looking for and making use of structure to do algebra. 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 (see Section 7.2 for more information). The practices emphasized in the SAT Suite are central to the demands of postsecondary work. Problem solving requires test takers to make sense of problems and persevere to solve them, a skill highly rated by postsecondary instructors (Conley, Drummond, de Gonzalez, Rooseboom, & Stout, 2011). Modeling stresses applications characteristic of the entire postsecondary curriculum. Students will be asked throughout high school, college, and careers to make choices about which tools to use in solving problems. Finally, structure is fundamental to algebra and to other more advanced mathematics.

As indicated in the test specifications, the Math Tests have two portions. One comprises questions for which students are allowed to use calculators to solve the problems. The other comprises questions for which students are not allowed to use calculators to solve the problems. The blueprint for each of these sections can be found as Table A-3.7 and Table A-3.8, respectively, in Appendix 3: Test Development Procedures.

The no calculator portion allows the SAT Suite to assess fluencies valued by postsecondary instructors and includes conceptual questions for which a calculator isn't helpful. Meanwhile, the calculator portion gives insight into test takers' capacity to use appropriate tools strategically. The calculator is a tool that students must use (or not use) judiciously.

The calculator portion of the test includes more complex modeling and reasoning questions to allow test takers to make computations more efficiently. However, this portion also includes questions in which the calculator could be a deterrent to expedience, thus

assessing the appropriate use of tools. For these types of questions, test takers who make use of structure or their ability to reason will reach the solution more rapidly than test takers who get bogged down using a calculator.

Detailed Description of the Content and Skills Measured by the SAT Suite Math Tests

The SAT Suite has been redesigned to better align to what research shows students need to know and be able to do in order to be prepared for college and careers. This goal has led to a more focused SAT Suite with a balance across fluency, conceptual understanding, and application. In these and other ways, such as embedding mathematical practices, the SAT Suite is also a good reflection of college- and-career-ready standards.

Heart of Algebra: Linear Equations and Functions

Algebra is the language of much of high school mathematics, and it's also an important prerequisite for advanced mathematics and postsecondary education in many subjects. The SAT Suite Math Tests focus strongly on algebra and recognize in particular the fundamentals of the subject that are the most essential for success in college and careers. Heart of Algebra assesses students' ability to analyze, fluently solve, and create linear equations and inequalities. Students are also expected to analyze and fluently solve equations and systems of equations using multiple techniques.

To assess full command of the material, these problems vary significantly in form and appearance. Problems may be straightforward fluency exercises or may pose challenges of strategy or understanding, such as interpreting the interplay between graphical and algebraic representations or solving as a process of reasoning. Test takers are required to demonstrate both procedural skill and a deeper understanding of the concepts that undergird linear equations and functions to successfully exhibit a command of the Heart of Algebra.

Mastering linear equations and functions has clear benefits to students. The ability to use linear equations to model scenarios and to represent unknown quantities is powerful across the curriculum in the postsecondary classroom as well as in the workplace. Further, linear equations and functions remain the bedrock upon which much of advanced mathematics is built. Consider, for example, that derivatives in calculus are used to approximate curves by straight lines and to approximate nonlinear functions by linear ones. Without a strong foundation in the core of algebra, much of this advanced work remains inaccessible.

For the content dimensions within the Heart of Algebra domain, see Table A-3.9 in Appendix 3: Test Development Procedures.

Problem Solving and Data Analysis: Proportional Relationships, Percentages, Complex Measurements, and Data Interpretation and Synthesis

The SAT Suite Math Tests have responded to the research evidence identifying what is essential for college readiness and success by focusing significantly on problem solving and data analysis: the ability 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. Problems in this category require significant quantitative reasoning about ratios, rates, and proportional relationships and place a premium on understanding and applying unit rate.

Interpreting and synthesizing data are widely applicable skills in postsecondary education and careers. In the SAT Suite Math Tests, test takers are expected to identify quantitative

measures of center, the overall pattern, and any striking deviations from the overall pattern and spread in one or two different data sets. This includes recognizing the effects of outliers on the measures of center of a data set. In keeping with the need to stress widely applicable prerequisites, the SAT Suite emphasizes applying core concepts and methods of statistics rather than broadly covering a vast range of statistical techniques.

Finally, the SAT Suite Math Tests emphasize the ability of test takers to apply math to solve problems in rich and varied contexts and feature problems that require the application of problem solving and data analysis to solve problems in science, social studies, and career-related contexts.

For the content dimensions within the Problem Solving and Data Analysis domain, see Table A-3.10 in Appendix 3: Test Development Procedures.

Passport to Advanced Math: Analyzing Advanced Expressions

As a series of assessments that leads to an entry point to postsecondary education and careers, the SAT Suite Math Tests include topics that are central to the ability of students to progress to later, more advanced mathematics. The problems in Passport to Advanced Math cover topics that have great relevance and utility for college and career work.

Chief among these topics is the understanding of the structure of expressions and the ability to analyze, manipulate, and rewrite these expressions. This includes an understanding of the key parts of expressions, such as terms, factors, and coefficients, and the ability to interpret complicated expressions made up of these components. Test takers must also be able to show their skill in rewriting expressions, identifying equivalent forms of expressions, and understanding the purpose of different forms.

This category also includes reasoning with more complex equations, including solving quadratic and higher-order equations in one variable and understanding the graphs of quadratic and higher-order functions. Finally, this category includes the ability to interpret and build functions, another skill crucial for success in later mathematics and scientific fields.

For the content dimensions within the Passport to Advanced Math domain, see Table A-3.11 in Appendix 3: Test Development Procedures.

Additional Topics in Math

Although the overwhelming majority of problems on the SAT Suite Math Tests fall into the first three domains, the SAT test and the PSAT/NMSQT and PSAT 10 tests also address additional topics in high school math. Patterns of selection for these are governed by evidence about their relevance to postsecondary education and work. The additional topics include essential geometric and trigonometric concepts and the Pythagorean theorem, which become powerful methods of analysis and problem solving when connected to other math domains.

For the content dimensions within the Additional Topics in Math domain, see Table A-3.12 in Appendix 3: Test Development Procedures.

Content Specifications for the SAT Essay

This section describes the content, format, and distinctive new features of the SAT Essay (found only on the SAT) as well as the skills it measures.

Overall Claim for the Essay

As with the other tests in the SAT Suite of Assessments, the optional 50-minute SAT Essay is intended to collect evidence in support of a broad claim about student performance:

> Students can demonstrate college and career readiness proficiency in producing a cogent and clear written analysis using evidence drawn from an appropriately challenging source text written for a broad audience (College Board, 2014, p. 69).

Test Description

While the above-mentioned source text varies from administration to administration, the Essay prompt itself is highly consistent. Such transparent consistency allows test takers, in their preparation and during the actual test, to focus squarely on source analysis and use of evidence in the specific text they are to analyze.

All passages are taken from high-quality, previously published sources. While the specific style and content of the passages inevitably vary to some extent, given the College Board's commitment to using authentic texts with this task, the passages take the general form of what might be called arguments written for a broad audience. That is, the passages examine ideas, debates, trends, and the like in the arts, the sciences, and civic, cultural, and political life that have wide interest, relevance, and accessibility to a general readership. Passages tend not to be simple pro/con debates on issues but instead strive to convey nuanced views on complex subjects. They are notable, too, for their use of evidence, logical reasoning, and/or stylistic and persuasive elements. Text complexity of the passages is carefully monitored to ensure that the reading challenge is appropriate and comparable across administrations but not an insurmountable barrier to test takers responding to the source text under timed conditions. Prior knowledge of the passages' topics isn't expected or required.

For the SAT Essay, test takers are asked to explain how the author of the accompanying passage builds an argument to persuade an audience. Test takers are informed that they may analyze such aspects of the passage as the author's use of evidence, reasoning, and stylistic and persuasive elements but that they may also or instead choose other features to analyze. They are further advised that, in all cases, they should center their discussion on those aspects that are most salient to the passage in question. Responses aren't to focus on whether test takers agree or disagree with the claims made in the passage but rather on how the author builds an argument to persuade an audience. In broad terms, responses are evaluated for demonstrated comprehension of the source text, the quality of analysis of that source text, and the quality of the writing in the response. Test takers' responses should demonstrate such dimensions as a careful understanding of the passage; effective, selective use of textual evidence to develop and support points; clear organization and expression of ideas; and a command of the conventions of standard written English. A fuller list of criteria used to evaluate responses is provided in Appendix 3: Test Development Procedures.

In a break from the past and the present of many standardized direct-writing assessments, the SAT Essay task is not designed to elicit test takers' subjective opinions but rather to assess whether they are able to comprehend an appropriately challenging source text and to craft an effective written analysis of that text. Rather than merely asking test takers to emulate the form of evidence used by drawing on, say, their own experiences or imaginations, the Essay requires them to make purposeful, substantive use of textual

evidence in a way that can be evaluated objectively. The Essay also connects reading and writing in a manner that both embodies and reinforces the interdependency of these literacy skills. Considered together with the MC SAT Reading and SAT Writing and Language Tests, the Essay response gives rich, detailed insight into students' reading and writing achievement and their readiness for college and careers.

Although the College Board remains steadfast in its commitment to the importance of analytic writing for all students, two factors have contributed to its decision to no longer make the Essay a required part of the SAT. First, while the writing work that students do in the Evidence-Based Reading and Writing section of the assessment is strongly predictive of college and career readiness and success, one single essay historically hasn't contributed significantly to the overall predictive power of the test. Second, feedback from hundreds of member admission officers was divided: Some of them found the Essay useful, but many did not. Therefore, by making the Essay optional, colleges have the flexibility to make their own decisions about requiring the Essay, and students applying to colleges that don't require the Essay are saved the expense and time for test results that won't be considered.

Test Summary

For the full content specifications of the SAT Essay and a list of the dimensions of the SAT Essay domain, see Table A-3.13 and Table A-3.14, respectively, in Appendix 3: Test Development Procedures.

Key Features of the SAT Essay

Three distinctive features of the SAT Essay are described below:

- Use of a common prompt
- Emphasis on analysis of argument
- Use of clear, powerful evaluation criteria

Common Prompt

In the Essay, test takers are asked to write a cogent and clear response based on the comprehension and analysis of a source text, supporting their claims and points about the text with evidence drawn from the passage. While the source text is different for each form of the SAT, the prompt is largely consistent in format and wording across administrations, as shown in Figure 3.1.

Because the prompt is largely the same from test administration to test administration, test takers can prepare by developing the underlying reading, analysis, and writing skills measured on the test rather than trying to anticipate the kind of question that will be asked. Moreover, because the Essay task is centered on a unique source text disclosed only on test day, students must engage with the passage rather than rely on canned, generic responses generated ahead of time. In these ways, the test encourages meaningful practice aligned with curriculum and instruction rather than narrow "prep" focused on mastery of an artificial test format.

Analysis of Arguments

The Essay requires students to analyze how an author uses evidence, reasoning, and/or stylistic or persuasive elements (and/or other elements of the students' choosing) to build their argument.

Figure 3.1: SAT Essay prompt

As you read the passage below, consider how [the author] uses:

- **evidence, such as facts or examples, to support claims;**
- **reasoning to develop ideas and to connect claims and evidence; and**
- **stylistic or persuasive elements, such as word choice or appeals to emotion, to add power to the ideas expressed.**

[Source Text]

Write an essay in which you explain how [the author] builds an argument to persuade [his/her] audience that [author's claim]. In your essay, analyze how [the author] uses one or more of the features listed above (or features of your own choice) to strengthen the logic and persuasiveness of [his/her] argument. Be sure that your analysis focuses on the most relevant aspects of the passage.

Your essay should not explain whether you agree with [the author's] claims but rather explain how the author builds an argument to persuade [his/her] audience.

Evaluation Criteria

The criteria by which students' written responses are evaluated are notable for their clarity and robustness. Each response is assessed using three analytic traits—Reading, Analysis, and Writing—each of which is scored on a scale of 1–4 (see Table A-3.15 in Appendix 3: Test Development Procedures).

These criteria are both aligned with and supportive of important priorities in rigorous high school instruction. The clarity and richness of the criteria yield important information about student performance that can be easily understood and translated into further classroom-based work and support.

3.2 Item Development for the SAT Suite of Assessments Reading, Writing and Language, and Math Tests

Item Specifications

All items on the SAT Suite Reading Tests and Writing and Language Tests, and most of the items on the SAT Suite Math Tests, are four-option, MC items, with each item having one and only one correct or best answer. Some items on the Math Tests are student-produced response (SPR) items, which require the student to solve a problem and then grid their response on the answer sheet. As stated earlier in this chapter, the prompt for the SAT Essay is largely the same in format and wording across administrations, with the source text being different from test form to test form.

In keeping with Standard 4.7 (AERA/APA/NCME, 2014), the following section describes how the College Board creates and reviews the MC and SPR items that comprise the tests in the SAT Suite of Assessments.

Crafting of Items

Test and Question/Task Specifications

Given the defined test domains, the College Board measurement and content staff worked with educational experts to prepare test and question/task specifications that represent the depth and breadth of the defined domains and to help ensure the consistent development of assessments of the highest quality. The specifications define the question/task types and formats required to measure most directly and authentically the domains of knowledge, skills, and understandings relevant to the SAT Suite of Assessments' primary purpose and the tests' overall claims.

Passage Selection and Question/Task Design

The SAT Suite measures durably powerful knowledge, skills, and understandings needed in postsecondary education, work, and life. All content area tests are developed to elicit from students work worth doing through questions and tasks that resemble the best classroom practices. The College Board does this by working with a multitude of teachers in K–12 and postsecondary instructors of entry-level courses across the U.S.

Reading Test

In the SAT Suite Reading Tests, students engage with texts worth reading and worthy of careful consideration. All passages are selected from previously published authentic writing that exemplifies the genres represented on the test and are designed to be powerful, insightful pieces. The essential first step of question development is a close and careful reading of the text.

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. They also favor a more organic development process that respects the unique natures of rich, authentic texts in a variety of content areas. In addition, questions on the SAT Suite Reading Tests unfold in a thoughtful sequence that helps make the investigation of the passage more natural and meaningful for students.

Writing and Language Tests

The SAT Suite Writing and Language Tests comprise passages that are engaging and challenging, paired with questions that focus clearly on a core of writing and language skills empirically linked to college and career readiness requirements. These commissioned passages (passages written for the test rather than excerpted or adapted from preexisting sources) are designed to provide meaningful contexts for the skills being addressed and exemplify the qualities of effective arguments, informative/explanatory texts, and nonfiction narratives. Test questions assess writing and language skills and understandings in extended prose contexts rather than in isolation and require students to make active choices in revising and editing rather than simply identify errors.

Math Tests

The Math Tests ask students to demonstrate their command of the mathematics most provably useful in a range of college courses and career environments. The SAT Suite Math Tests provide the opportunity for richer applications of the most essential math to address real-world situations and problems and include multipart applications of this core of useful math. These core topics are examined extensively and at a very high level of proficiency.

Test questions are thoughtfully designed with the help of teachers with a deep knowledge of the target mathematical content and practices. The questions on each Math Test explore the full dynamic range of each content area through precisely crafted questions that emphasize the use of math in unlocking insights and solving problems. The test design allows the core of math to be examined with the range of rigor required (as defined through evidence) to be college and career ready, examining at once students' procedural skill, application, and understanding. Rather than covering a broad number of topics that most students will never see again, students study fewer topics that represent a deep core that they can draw upon again and again in their schooling, college, and career. At the same time, the assessment includes pure math problems that focus on abstract reasoning essential for success in solving diverse problems and engaging in demanding disciplines.

Question/Task Writer Support Materials

In order to consistently develop assessments with engaging, authentic stimulus materials and contexts that lend themselves to high-quality questions and tasks, the College Board has developed and continues to maintain a range of test-support materials intended to help make sure that all questions and tasks are evidence based, valid, and accessible to all students—in short, that they meet the highest possible standards (AERA/APA/NCME, 2014). These materials include question/task writer content and fairness guidelines as well as question/task prototypes and templates. The College Board contracts with classroom teachers at both the high school and postsecondary levels and with other independent content and instructional experts to develop and/or review all questions and tasks. In this way, those most familiar with the student population of interest and knowledgeable in the instructional best practices in the field make the most significant contribution to assessment content. This helps ensure that the test materials and tasks included in the assessments are engaging, instructionally appropriate, and fair to all students.

Item Content and Fairness Reviews

Prior to pretesting, all questions and tasks are reviewed by external, independent reviewers who are asked to evaluate each question and task according to a set of criteria for content accuracy and fairness. In keeping with Standard 4.8, these reviewers are typically active classroom teachers drawn from across the nation from both the secondary and postsecondary levels and are deeply familiar with the student population of interest and the nature and purpose of the test (AERA/APA/NCME, 2014).

Content reviewers are focused on ensuring the soundness of each question, task, and stimulus and evaluating its relationship to the construct (e.g., reading) being measured, its relevance and appropriateness to the work students do in high school, and its value in terms of measuring students' degree of college and career readiness. Fairness reviewers are charged with helping ensure that test questions, tasks, and stimuli are broadly accessible to the wide-ranging student population that takes the assessment, that the questions are clearly stated and unambiguous in their intent, and that the questions and tasks don't offer unfair advantages to some students. See Section 2.2 for more information on the fairness review.

Item Pretesting and Analysis

Item Pretesting

Every item used in an operational form of the SAT Suite of Assessments has previously been pretested; that is, the item is tried out with an appropriate group of test takers to make sure that it isn't ambiguous or confusing and to determine the difficulty level and the degree to which it differentiates among higher- and lower-achieving test takers (AERA/APA/NCME, 2014). The pretest responses are also analyzed to determine whether test takers of different racial/ethnic groups or gender groups respond to the question differently.

SAT Suite item writing and review are ongoing activities throughout the year. The MC items in Reading, Writing and Language, and Math, as well as the SPR items in Math are pretested on a motivated sample of students that resembles the SAT Suite population of interest. Pretests are assembled from questions that have received a number of content, fairness, and editorial reviews prior to pretesting. Pretests are administered to students in test administrations like those in which the SAT Suite of Assessments is given. The data from 1,000 to 3,000 students responding to each question are used to evaluate question performance. This item information provides an accurate estimate of how the item will function when administered as part of a future SAT, PSAT/NMSQT and PSAT 10, or PSAT 8/9.

Analysis of Pretest Information

In keeping with Standard 4.10, data collected from MC and SPR pretests are analyzed to provide important information about the appropriateness of items for use in operational forms of the SAT Suite (AERA/APA/NCME, 2014). Three statistical indices are computed: equated *p*-value as an index of item difficulty within the SAT Suite population, r-biserial as an index of whether the item discriminates between higher- and lower-achieving test takers, and Mantel-Haenszel DIF (differential item functioning) as an index of the relationship between group membership and the likelihood of answering the question correctly. These item statistics are used to judge whether a given question is suitable for inclusion in the pool of items from which operational forms are assembled. The item statistics may also reveal problems with the conceptualization or wording of a question. Some of these items will be revised and re-pretested. Others will be discarded.

Each SAT Essay passage is reviewed by College Board staff and by members of the SAT Writing and Language Test development committee. After all content and fairness concerns raised during the review process are resolved, the passage and prompt are pretested in a special administration in high school English classrooms. For each group of pretests, a diverse sample of schools is invited to participate by having students respond to a particular passage during their English class. A sample of at least 300 responses to each pretest passage is obtained in order to provide a wide range of essays for determining whether the passage is accessible to students and to provide exemplars of various levels of reading, analysis, and writing skill for use in training readers.

The responses to each pretested Essay passage are scored by a group of experienced readers who are trained to apply the SAT reading, analysis, and writing rubrics using exemplar essays gleaned from the pretest samples. For each pretested passage, the distribution of reader-assigned scores is analyzed to determine whether high school students readily understand the passage and whether it elicits the full range of student

achievement in reading, analysis, and writing. In other words, does the passage lead to responses that can be scored reliably and that provide differentiation among higher- and lower-achieving students in all three domains? A passage that is determined not to elicit the full range of responses, or that elicits responses that are markedly skewed toward either end of that range, are deemed not usable for operational testing. In addition to the full distribution of essay scores, the distributions of scores by student gender or ethnicity are compared in order to uncover those passages that appear to favor one demographic group over another. Passages that appear to do this are also deemed unusable for operational testing.

SAT Suite Item Characteristics

The statistical indices employed in analyzing and screening SAT Suite items are as follows.

Item Difficulty

The difficulty of an item is a function of the percentage of test takers who answer it correctly (i.e., *p*-value). An item's difficulty should be appropriate for the population taking the test. When an item is too easy, virtually all test takers answer it correctly; thus, extremely easy items contribute very little information to the total test score. Similarly, inappropriately difficult items aren't very useful in a test. Because items within a test are highly intercorrelated, it is best to select items with a moderate spread of difficulty around a mean *p*-value of .5 (or 50% correct).

Because the samples to which specific pretest items are administered may, to some degree, differ in achievement level from test date to test date, it's necessary to convert the raw *p*-values to equated *p*-values. To make this conversion, items with known equated *p*-values are administered along with the pretest items. The raw *p*-value for these items based on the pretest sample is then transformed into a delta and plotted against the known equated deltas (i.e., transformed *p*-values) for these items. The resulting linear relationship between the pairs of raw and equated deltas is used to compute an equated delta and *p*-value for the new pretest items (See Exhibit 5.1 in Appendix 5: Interpretation and Application of Results). An equated *p*-value is computed for each pretest item and is based on the standard reference population, permitting comparisons of items across samples.

Item Discrimination

Although difficulty level is one important criterion in selecting items, item discrimination is essential to be able to distinguish among test takers at different levels of achievement. The r-biserial correlation coefficient between the item and the total test score is most often used to assess the item's utility in discriminating among test takers of differing achievement levels and the homogeneity of test items (or extent that test takers' performance on an item relates to their total test score). The biserial correlation ranges from +1 to -1. The more positive the correlation, the more the item distinguishes test takers with high total scores from those with low scores. A negative biserial correlation indicates that the item is measuring something different from the rest of the test: test takers with high scores are more likely to answer that item incorrectly than those with low scores. Correlations that are near zero indicate that high scorers and low scorers have the same chance of correctly answering the item. Therefore, the SAT Suite doesn't include items with low or negative biserial correlations.

Biserial correlations also provide an indication of the homogeneity of test items. If the correlation is very close to +1, all of the information provided by the item is redundant with that provided by the other test items. Items with moderate biserial correlations distinguish among ability levels, yet also supply unique information. Therefore, most items included on SAT Suite operational forms fall within a biserial range of +.20 to +.80. See Exhibit A-5.1 in Appendix 5 for the formula for r-biserial correlation.

In determining whether to select, omit, or edit and refine a multiple-choice item based on results from pretests, test developers also consider the number and percentage of test takers who respond to the correct option and to each incorrect option. At each score level for the total multiple-choice test, the percentage of test takers selecting each option of the item is plotted. For the correct option of the item, it's expected that the percentage of test takers selecting that option increases as the total test score increases. An example plot for a well-functioning item where the percentage of test takers selecting the correct option increases with the total multiple-choice score is shown as Figure A-3.1 in Appendix 3: Test Development Procedures, where option A is the correct option and its series increases with the raw score. If the correct option doesn't display this pattern, the item is carefully reviewed. Similarly, if an incorrect option has this typical increasing pattern, then that option is closely evaluated. As a result of the evaluation, the item may be revised and then re-pretested, or it may be discarded entirely.

Differential Item Functioning

Analyses of differential item functioning (DIF) are conducted to identify items that may function differently for members of different groups. DIF analyses compare the performance of two groups of test takers (e.g., males vs. females, Asian American test takers vs. White test takers) who have been matched on their reading, writing and language, or math achievement (SAT Suite Reading, Writing and Language, or Math test score) on each item. 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 item correctly, regardless of gender, race, or ethnicity. DIF occurs when individuals with similar scores on the Reading Test, the Writing and Language Test, or the Math Test differ notably in their performance on a specific test item. The presence of DIF indicates that an item functions differently for one subgroup from the way it functions for another subgroup of the same achievement level.

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 (AIAN), Asian, Black or African American, Multiple Races, and White. All non-Hispanic respondents are identified as one of the previously listed racial categories. At the test development stage, the minimum sample size requirement for the focal group is 100 when calculating the statistics. During operational administration analyses, the minimum sample size requirement for the focal group is 200.

The Mantel-Haenszel (MH) procedure (Dorans & Holland, 1993) is used for DIF analyses with the SAT Suite. For the formulas involved in the MH DIF statistical procedure, see Exhibit A-5.1 in Appendix 5. This procedure computes a ratio for the conditional probability of successful reference group performance on an item over the conditional probability of successful focal

group performance on the item for each score level on the test. Thus, comparisons are made of test takers with equivalent scores (e.g., equivalent achievement in mathematical reasoning) at each point on the test. Statistically optimal weights are then assigned to each ratio, and they are averaged across all score points. The MH statistic is transformed to a delta scale, and the resulting statistic is referred to as the Mantel-Haenszel delta DIF (MH D-DIF). The MH D-DIF statistic ranges from negative infinity to infinity, with a value of 0 indicating no DIF. Both the magnitude of the MH D-DIF and a significance test are used to evaluate the presence or absence of DIF. For the SAT Suite, MH D-DIF values are considered:

- negligible if they are between 1.0 and -1.0 or aren't statistically different from zero at the .05 significance level;
- moderate if they fall between 1.0 and 1.5 or -1.0 and -1.5, or if they're greater than 1.5 or -1.5 and not significantly different from the absolute value of 1.0 at the .05 significance level; and
- sizable if they exceed 1.5 or -1.5 and are statistically different from the absolute value of 1.0 at the .05 significance level.

Items having sizable values of DIF, those items favoring one group over another for test takers of the same level of achievement undergo further review to determine whether some aspect of what the item is measuring is particularly related to subgroup membership and irrelevant to the dimension being measured. When an item is identified as exhibiting such characteristics, it is either revised and re-pretested or eliminated. Items exhibiting moderate DIF may be selected for a final form if items with negligible DIF are insufficient to meet particular specifications.

Statistical Specifications

In addition to the content specifications, each form within each of the SAT Suite of Assessments is built to a common set of statistical specifications. This is to ensure that each form is of comparable difficulty, discrimination, and score reliability to every other form for that specific assessment, in addition to being of comparable content. Together, this content and statistical comparability is necessary to ensure that each new form of the SAT, PSAT/NMSQT and PSAT 10, and PSAT8/9 may be equated to previous forms. Equating, in turn, ensures that the scale scores and subscores reported to students have the same meaning, in terms of achievement, regardless of which form a student takes (for more information on Equating, see Section 6.2).

A total of 11 or 12 equated scale scores and subscores are reported for the SAT Suite of Assessments:

- The Reading, Writing and Language, and Math Test scores
- The Analysis in Science and Analysis in History/Social Studies cross-test scores
- The Expression of Ideas, Standard English Conventions, Words in Context, Command of Evidence, Heart of Algebra, Problem Solving and Data Analysis, and Passport to Advanced Math (SAT and PSAT/NMSQT and PSAT 10 only) subscores

All forms within each test in the SAT Suite of Assessments are constructed to be as similar as possible to one another, in terms of item difficulty, item discrimination, and score reliability, along all 12 of these score and subscore areas. Item difficulty is estimated by the item's *p*-value. Item discrimination is estimated by the item-total biserial correlation coefficient. Score reliability is estimated by the Kuder-Richardson (KR20) coefficient.

3.3 Test Form Assembly

Item Bank

Once items are developed and pretested on a representative sample of the population of interest as discussed in Section 3.2, they are officially stored and maintained in the College Board's item bank. The item bank contains all items developed for each test of the SAT Suite coded by their item classifications, as specified by the test specifications. For all passages, items, and stimuli, the bank holds the text, art/graphics codes, item codes for required elements, and statistical records for appropriate test administrations. Information held in the item bank also allows the appropriate items to be selected and placed into the test assembly item pool for assembling into test forms.

Initial Test Form Assembly

Multiple SAT Suite operational test forms are developed each year to ensure that all administrations are covered with the needed number of secure forms. Every form is carefully assembled to meet the test specifications, both content specifications and statistical specifications, as discussed in Section 3.2. The content specifications are in Appendix 1 of this manual. A test form construction guide detailing the process and the established expectations is followed closely each time test forms are assembled. In following specified procedures and specifications, all forms of the test are created to be parallel to each other in the content that they cover; the knowledge, skills, and understandings that they assess; and the difficulty of the items. Although a statistical procedure called equating (see Section 6.2: Equating) is conducted for each test form to ensure that scores are interchangeable across multiple test forms and test dates, this procedure couldn't be effective if the test forms weren't already as parallel as possible in terms of content and difficulty.

Given the size of the item pool (thousands) used to construct test forms each year and the large number of forms (generally between 8 and 12) to be constructed at any one point in time, an automated test assembly (ATA) program is used to initially construct the forms to ensure that each and every test form is assembled to meet the test specifications as closely as possible. The ATA program contains all of the constraints that are specified in the test specifications and needed to meet statistical targets and to provide a list of the items chosen for each form. These initial test forms are then reviewed by measurement specialists to ensure that they meet all specifications before they are provided to the content experts for a thorough review of not only each and every item but also the test form as a whole.

Test Form Review

Content experts within the College Board review the test forms to ensure that the passages and/or items selected for each test form work well together from a content perspective. For example, items or passages from one part of the test shouldn't provide any clues to items on another part of the test. After several content experts within the College Board have reviewed and approved the initial test forms, they will be sent to external content and fairness reviewers. These reviewers are classroom teachers at either the secondary or postsecondary level, are experts in their field, and are familiar with the population of interest. Anywhere from 8 to 13 external experts review each and every test form. These experts review the form independently and then come together as a group with the other reviewers and with College Board content experts to discuss every item on the form, as well as the form as a whole. At this stage of the process, if the reviewers find any problems with any of the items or passages, the

problem items or passage sets will be replaced with items and/or passages from the remaining pool. All replacement items and passages result in a new review cycle by the measurement specialists and the content experts (both internal and external to the College Board). This review process continues in an iterative fashion until all test forms are given final approval.

Once the forms receive final approval by both the measurement specialists and the content specialists, they are proofed and prepared for publishing.

3.4 Passage and Prompt Development for the SAT Essay

The College Board has developed a rich task for the SAT Essay that authentically reflects the work students need to do to be ready for and successful in college and careers. The SAT Essay is optional, with the decision whether to consider it as an admission requirement being at the discretion of each institution of higher education.

As stated in Section 1.3, the basic aim of the SAT Essay is to determine whether students can demonstrate college and career readiness proficiency in reading, writing, and analysis by comprehending a high-quality source passage and producing a clear and cogent written analysis of that passage, supported by critical reasoning and evidence drawn from the source. Although the Essay source passage varies from administration to administration, the Essay task itself and the prompt language is highly consistent. Such transparent consistency allows students, in their preparation and during the actual test, to focus squarely on source analysis and use of evidence in the specific texts they will analyze. See Section 3.1 for more information on the Specifications of the Essay Test.

Crafting of Passages and Prompts

As described in Section 3.1, all Essay passages are taken from high-quality, previously published sources. While the specific style and content of the passages vary, the passages take the general form of what might be called arguments written for a broad audience. The passages examine ideas, debates, trends, and the like in the arts, the sciences, and civic, cultural, and political life that have wide interest, relevance, and accessibility to a general readership. Essay passages are also selected based on their use of evidence, logical reasoning, and/or stylistic or persuasive elements so that students have the best opportunity to demonstrate their ability to analyze how the author(s) built the argument. Additionally, the text complexity of the passages is carefully monitored to ensure that the reading challenge is appropriate and comparable across administrations, but not an insurmountable barrier to test takers responding to the source text under timed conditions. Prior knowledge of the topics of the passages is not expected or required.

As much as is possible, passages used for the SAT Essay are kept intact as they originally appeared in publication. In other words, the College Board avoids, or limits as much as possible, the editing or excising of portions of a passage. This allows the passage to maintain the integrity of the argument as the author(s) originally wrote it. Only very minor and limited edits or deletions are made for the purpose of word length, fairness, sensitivity, or obscurity/difficulty. When elisions or minor additions are made, they're indicated by ellipses or brackets in order to maintain transparency for readers. Any additional editing or adaptation beyond minor changes risks further compromising the passage as it originally appeared.

Essay passage finders are trained to locate passages that are suitable for use on the SAT Essay according to the criteria articulated earlier in this chapter. As part of the passage

finding process, the College Board test development staff selects passages, which are then either accepted or rejected for further development based on discussions and evaluations by the full College Board Essay development team.

Once Essay passages have been accepted, the College Board crafts a prompt-specific summary statement for each passage. This statement comes at the end of each passage in the prompt text box that outlines the Essay task. The prompt summary statement provides test takers with the main claim of the passage so that they can focus on demonstrating a more detailed understanding of the passage and on analyzing how the author(s) built the argument.

Essay Passage Content and Fairness Reviews

As noted previously, Essay passages are selected based on their appropriateness and suitability for a wide audience. Once passages have met internal criteria for selection based on topic, length, and text complexity, secondary and postsecondary classroom teachers then review the prompts for any issues of fairness. In keeping with AERA/APA/NCME Standard 3.2, the test developers work to ensure that scores aren't influenced by construct-irrelevant characteristics (AERA/APA/NCME, 2014). Prompts are reviewed to ensure that they don't disadvantage any particular subgroup in the student population based on factors such as race, ethnicity, gender, sexual orientation, or socioeconomic status. Because an argument by its very nature must be debatable to some degree, the subject matter of SAT Essay passages occasionally falls into potentially sensitive territory. Topics for the SAT Essay may sometimes be perceived as sensitive because they encompass rich and complex issues in which differing perspectives may sometimes come into conflict. However, the College Board is careful to select topics that are appropriately subject to debate and yet unlikely to cause an emotional response for students. For more information on Fairness Reviews prior to testing, see Section 2.2.

Essay Passage and Prompt Field Testing and Analysis

All SAT Essay passages and prompts are put through rigorous field testing to ensure the comparability of passages across student populations.[1] The students who participate in field testing vary by race, ethnicity, and socioeconomic status. They come from rural, suburban, and urban populations, and the field test population includes both private and public school students. In keeping with AERA/APA/NCME Standard 4.8, it is important to ensure that the prompts "function similarly for different groups" (AERA/APA/NCME, 2014, p. 88). Once a sufficient testing volume is reached to enable the analysis of the performance of each passage and prompt, the College Board scores these essays according to the SAT Essay rubric (see Table A-5.1 in Appendix 5: Interpretation and Application of Results). For more information on SAT Essay scoring, see Section 5.1. The procedures used to score the field test essay are the same used to score the Essay after it is administered to test takers. Each essay is scored independently by two raters who assign a score of 1–4 in each of three dimensions: Reading, Analysis, and Writing. If the two independent rater scores are in perfect agreement or are adjacent to each other, the scores from the two independent raters are added together to give test takers a 2–8 score on each of the three dimensions. If the two independent raters scores are nonadjacent scores (e.g., scores of 1 and 4), the SAT Essay

[1] This is a different kind of comparability than that related to Test Accommodations, which was discussed in Section 2.3.

response is rated by a third senior-level rater. The third rater's score is then doubled to give test takers a 2–8 score on the relevant dimension(s).

To ensure fair and comparable SAT Essay prompts, the College Board completes psychometric analyses on the field test results of all prompts, including interrater reliability stats (including correlations) between Rater 1 and Rater 2 (and Rater 3 if applicable) for all dimensions (Reading, Analysis, Writing). For information on item pretesting and analysis for MC and SPR items, see Section 3.2. The mean and the standard deviation for each prompt are compared, as well as the frequency distribution by score point across each dimension for each prompt. Any prompt that is an outlier—that has a mean score on any dimension that is unusually high or low or that has a distribution of scores on any dimension that is significantly different from the other prompt distribution of scores—isn't used operationally. Demographic analyses are also conducted, including analyzing female/male mean differences by dimension scores by prompt and mean differences by race for dimension scores by prompt. Any prompt that demonstrates extreme differential performance, either between males and females or between White and other races, isn't used operationally.

3.5 Accommodated Forms

The College Board believes that, in keeping with AERA/APA/NCME Standards, "All test takers should have the full opportunity to demonstrate their standing on the construct being measured" (AERA/APA/NCME, 2014, p. 52). Consistent with the Americans with Disabilities Act, and to ensure fairness across assessments, the College Board's policies and procedures are designed to ensure that appropriate testing accommodations are made available to test takers with disabilities. The College Board's Services for Students with Disabilities (SSD) office authorizes a broad range of test accommodations. These accommodations include, but aren't limited to, braille tests, large print tests, and extended testing time. Students who show that their disabilities affect their ability to participate in the SAT Suite of Assessments are eligible for accommodations.

Available since 2010, the online request processing system (www.collegeboard.org/students-with-disabilities) allows schools to request accommodations and allows parents and schools to track the progress of the request. Parents may request accommodations without the participation of schools via a paper request form. Once approved for accommodations by the SSD office, students are permitted those accommodations on the SAT, PSAT/NMSQT, and PSAT 10 assessments. In addition to providing a fair test-taking experience to all users of the assessment, it's also intended that these accommodations respond to any test taker characteristics that could potentially interfere with the validity of test score interpretation (AERA/APA/NCME, 2014).

The following examples of accommodations available from the College Board ensure that eligible students receive the accommodations they need. In keeping with AERA/APA/NCME Standards, these accommodations yield inferences that are comparable to those from the standard version of the assessment (AERA/APA/NCME, 2014). Many accommodations are administered in the standard testing room, as part of the administration. In select instances, an SSD coordinator administers accommodated tests in a separate space. Schools don't need to request approval from the College Board to administer accommodations on the PSAT 8/9 assessments at this time, but all of the accommodations listed in the following text are available for that assessment as well.

Please note the accommodations listed below are only examples—the list is not exhaustive.

Presentation

- Large print (14 pt., 20 pt.)
- Reader (Note: Reader reads entire test.)
- Use of a highlighter
- Sign/orally present instructions
- Visual magnification
- Audio recording (MP3 audio)
- Colored overlays
- Braille
- Braille graphs
- Braille device for written responses
- Use of computer for essays (SAT only)
- Assistive technology compatible (screen reader accessible) format

Responding

- Verbal; dictated to scribe
- Computer without spell check/grammar/cut and paste features (SAT only)
- Record answers in test booklet
- Large block answer sheet
- Use of four-function calculator in Math Test – No Calculator

Timing and Scheduling

- Frequent breaks
- Extended time
- Multiple day (may or may not include extra time)
- Specified time of day

Setting

- Small group setting
- Private room
- Alternative testing site (with proctor present)
- Preferential seating

Supports for English Language Learners

To make the SAT even more accessible to students, the College Board has worked with educators and state partners over the past year to provide testing supports for English language learners (ELL).

Effective Jan. 1, 2017, ELL students taking a state-funded SAT during the school day have access to testing instructions in several native languages[2] and approved word-to-word bilingual glossaries.

ELL students taking a state-funded SAT during the school day don't have to apply to use word-to-word bilingual glossaries or translated instructions. Schools will be able to administer them directly to their students as needed, and students will receive a college-reportable score.

3.6 Test Form Production

Test form production consists of two stages: certification and manufacturing (printing/distribution).

Test Form Certification

After completing the rigorous quality development and approval process to construct the test forms, and before test forms are released for manufacturing, the College Board staff conducts a thorough content and editorial review of the test forms and their metadata to ensure test form quality. This includes a post-committee content review as well as a copyediting review, a key check review, a test layout review, and finally a cold read review before release to manufacturing.

The following descriptions outline the *high-level quality control* of test forms during the composition and layout stages.

Composition and Layout (prior to release for manufacturing)

Content Review. This is a thorough content review by a trained test developer other than the test assembler, as well as a copyediting review. The review checks for both item-level and test-level issues of any kind.

Key Check Review. Content staff performs an independent review of the correct answers/keys, essentially taking the test as a student might, to confirm that the keys are correctly represented in the metadata.

Layout Review. A review by a trained editor, who checks for issues such as unclear directions and formatting, presentation of items, general appearance, and various specific issues/concerns related to each item type in the form.

Cold Read. A proofreading (non-content) review performed on the certified test section by an editor who hasn't previously viewed the section, looking for any observable errors with "fresh eyes."

Certification by Test Section. This is a final review by the test assembler, who confirms the results of the layout review. The output of this review is certified copy of each page of a test section, followed by a coordination review.

Final Read of Complete Test Form. After a test form has been assembled, an end-to-end final read through of all item text, as well as a check of all collateral/template elements, is performed by an editor.

[2] At this time, languages covered by these supports include Spanish, Arabic, Portuguese, Polish, Mandarin, Haitian Creole, Russian, and Vietnamese.

In the event a late-stage change is identified during these processes, the item must receive a review by a second test developer, plus an additional editorial review and managerial approval of the change.

Test Form Manufacturing

Qualifications. Manufacturers that are selected to print test books have successfully demonstrated quality capability against formal audit criteria and are continuously monitored to ensure ongoing conformance.

Quality Standards. The College Board Print Quality Standards define the acceptable quality level (AQL) for exam books and supporting operational publications. In addition, expected content accuracy and final print quality are described in several vendor Service Level Agreements.

- Guidelines are provided on critical supports that are necessary for high quality, such as employee training programs, defect resolution, and corrective action.
- Defect classification is defined and aligned to the most stringent industry standards.
- Student interest is protected by the standard. A unique parameter for judging defects called "loss of information" is incorporated, stipulating that any anomaly that could render a test item incomplete or misleading is automatically classified as a major defect.

Control Plans. Print manufacturers are required to develop and maintain specific control points in their processes that support the print quality standards, thereby assuring that the printed output is of high quality directly off the manufacturing line, as opposed to applying extensive inspections post production.

Advanced Sample Review. A thorough review of a preproduction sample of the printed book is completed against a certified copy. This is a word-for-word check to ensure that the output of the printer's file is an exact match. An automated compare tool can be leveraged to scan and compare the proof against the certified copy.

Production Review. The printer performs a detailed inspection of the first production run of the book to confirm that the basic print characteristics match the certified copy and that the material meets the print quality standards. The printer conducts ongoing press pulls during the print run to verify consistency of print quality.

Materials Management Validation. Finished books are bundled and shipped according to specifications that have been verified to accurately reflect customer requirements. Warehouse inspections ensure that the right materials are designated to be shipped and that the content of each shipment is complete and accurate. Chain of custody is maintained during shipping of materials form the printer to the fulfillment warehouse.

CHAPTER 4

Testing Requirements

The 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 how we administer the SAT Suite of Assessments, as well as the steps taken to protect test materials and prohibit the inappropriate sharing of test information during the test administration. In keeping with best practices and the AERA/APA/NCME (2014) *Standards for Educational and Psychological Testing*, the College Board implements several steps to ensure these factors are taken into account during all phases of testing.

Section 4.1 discusses the policies and procedures involved in uniform test administration, an important component in providing a fair and equitable testing situation and ensuring that the scores produced from all administrations of an assessment are valid for their intended uses. We discuss the specific procedures undertaken to maintain uniform test administration and the rationale behind them. We also examine the roles and qualifications of the testing staff involved in the process.

Section 4.2 discusses some of the specifics of test security and the ways in which we prevent scenarios that would provide an unfair advantage and compromise the scores for their intended uses. We discuss these procedures as they apply to test materials and test takers, as well as the rationale behind these procedures.

4.1 Administration

In keeping with best practices and the AERA/APA/NCME Standards, the College Board has an established procedure that ensures the SAT Suite is administered to all test takers in a fair, equitable, and standardized manner (AERA/APA/NCME, 2014). The goal of the standardized administration process is to enable all test takers to experience a uniform set of conditions so that test scores from several different administrations can be used interchangeably for reporting, counseling students, and making admission and placement decisions.

The SAT is administered in two different models:

Weekend Administration. Test takers register to sit for the test at a nearby test center that may or may not be the school they regularly attend. They may register at any center with available seats. The administration takes place on an announced Saturday or Sunday, with Sunday administrations limited to test takers preapproved due to religious conflicts.

SAT School Day Administration. Test takers sit for the test at their school. The administration takes place on an announced school weekday.

The weekend administration of the SAT occurs seven times throughout the school year in the United States (and four times internationally) from August to June. The SAT School Day administration occurs once in October, twice in March, and twice in April, on a school weekday.

The PSAT-related assessments (PSAT/NMSQT, PSAT 10, and PSAT 8/9) are administered in a school-based model. Test takers sit for the test at their schools, either by electing to

participate or because they are in a cohort whose testing is sponsored by their school, district, or state.

The PSAT/NMSQT is administered on three announced days: two school weekdays and one Saturday. The administrations are typically held on (1) a Wednesday in mid-October, (2) the following Saturday, and (3) the last Wednesday in October. Schools select one administration for all of their participating students.

The PSAT 10 is administered during a spring test window, which typically begins in late February and ends in mid-to-late April.

The PSAT 8/9 is administered during two distinct test windows, one in the fall and one in the spring. The fall test window typically begins in late September and ends in January. The spring test window typically begins in late February and ends in mid-to-late April.

Schools select a PSAT 10 and/or a PSAT 8/9 administration date within the test window that is best for their scheduling needs.

Unless specifically noted otherwise, the procedures and policies described in this chapter apply to all administrations in the SAT Suite of Assessments.

Procedures

Highlights of the procedures and policies necessary to maintain test validity and fairness appear in the following text, but the most complete set of details can be found in the various administration manuals sent to the educators who administer the assessments.

The manual for SAT School Day can be found at **https://collegereadiness.collegeboard.org/sat/k12-educators/sat-school-day/downloads**.

Manuals for weekend SAT administrations can be found at **https://collegereadiness.collegeboard.org/sat/k12-educators/coordinating/testing-manuals**.

The manuals for PSAT/NMSQT and PSAT 10 can be found at **https://collegereadiness.collegeboard.org/psat-nmsqt-psat-10/k12-educators/resource-library**.

The manual for PSAT 8/9 can be found at **https://collegereadiness.collegeboard.org/psat-8-9/k12-educators/resource-library**.

Standardization

Uniform procedures are essential to a standardized testing program. The AERA/APA/NCME Standards state that "test administration conditions should be standardized for all examinees" (AERA/APA/NCME, 2014). By strictly following College Board policies and procedures, test center staff members provide test takers with a fair testing experience that ensures comparable and valid scores across all administrations of an assessment. These policies and procedures entail staff adhering to the same testing procedures and delivering instructions exactly as they appear in the test administration manuals.

In further keeping with AERA/APA/NCME Standards, the College Board is committed to test takers receiving comparable treatment during their test administration (AERA/APA/NCME, 2014). Even though the test forms or testing conditions might be slightly modified based on the needs of a particular student (e.g., with accommodations for students with disabilities), the construct being tested and the meaning of the score remain unchanged. No one is to suffer a disadvantage or gain an advantage of any kind because of race, religion, gender, or

disability. To maintain standardization across all administrations, all test takers are also to be protected from disturbance or other irregularities.

Testing Staff Qualifications

Each location where the SAT Suite is administered is supervised by an experienced educator provided with detailed instructions and scripts for administering the assessment in a uniform manner. The supervisor is responsible for all aspects of the test administration, including identifying staff who meet College Board qualifications, planning the use of facilities, and maintaining the security of test materials from the materials' arrival until their return. Qualified and competent test center staffers are integral to maintaining a fair testing experience and scores that are valid for intended uses across all administrations (AERA/APA/NCME, 2014). The test center staff should reflect the diversity of the test takers and are expected to act in a fair, courteous, nondiscriminatory, and professional manner. Prior to each test administration, the supervisor meets with staff to assign roles and responsibilities for the upcoming administration. As part of this preparation, the supervisor and staff review the rules and procedures involved in correctly administering the assessments, including those related to test security.

Associate (or room) supervisors and proctors assist the supervisor. The associate supervisor checks test taker identification, reads the test administration script verbatim, and manages all other aspects of the administration taking place in their assigned room. In rooms with more than 34 test takers, one or more proctors will assist the associate supervisor; the ratio is 1 proctor to every 50 test takers.

The staff members in each room are responsible for distributing and collecting test materials and telling test takers when to begin and end each test section. While students are working on the test, staff members walk around the room to guard against misconduct and make sure that each test taker is working on the appropriate section of the test and using appropriate pencils for marking the answer sheet. Staff members are also responsible for making sure that no test materials leave the room, and for reporting any irregularities to the supervisor.

Among other qualifications, test center staff should:

- Have unquestionable integrity and sound judgment.
- Be fluent in English and experienced in working with test takers.
- Not work for private test preparation for pay that is sponsored by non-school agencies or companies.
- Not have taken the assessments in the SAT Suite in the 180 days prior to administering the test.
- Have no family members taking the tests during the administration.
- Adhere to all published policies and procedures.

Detailed procedures and instructions for staff are provided in the test administration manuals.

The College Board offers a wide variety of professional development activities through in-person and technology-based options designed to prepare high school assessment coordinators, test administrators, and proctors for a successful test experience.

Mode

The SAT Suite is currently offered as a series of pencil and paper assessments. Digital versions of the assessments are under development, with select schools participating in pilot studies by invitation or under specific contractual agreements.

Timing

Each assessment in the SAT Suite is administered as three tests (Reading Test, Writing and Language Test, and Math Test). The SAT includes an optional Essay. The Math Test is divided into portions, one with calculator use allowed, and one without. The timing of each portion, including breaks, is listed in Tables 4.1 through 4.3. To ensure a standardized testing experience, staff should provide test takers with the appropriate amount of time in which to take the assessment—no more, no less. The chart doesn't include administrative activities such as distributing and collecting test materials and the test taker's completion of identifying information.

Test Materials

Test materials, including test books, answer sheets, and administration manuals are shipped via traceable courier to schools according to details gathered during the registration process. To maintain test security and expedite score reporting, SAT test materials are returned as soon as possible after the completion of testing. PSAT/NMSQT, PSAT 10, and most PSAT 8/9 test books are securely retained at the school to be given back to students when score reports become available. To prevent irregularities, test administrators are responsible for the receipt, distribution, and return of all relevant materials and for maintaining the security of these materials during the assessment process. We discuss security procedures in more detail in the following text.

Table 4.1: Timing of the SAT Assessment Including Breaks

Test	Time (in minutes)
Reading Test	65
Break	10
Writing and Language Test	35
Math Test – No Calculator	25
Break	5
Math Test – Calculator	55
Break (if taking Essay Test)	2
Essay Test (Optional)	50
TOTAL SAT/SAT with Essay	**195/247**

Table 4.2: Timing of the PSAT/NMSQT and PSAT 10 Assessments Including Breaks

Test	Time (in minutes)
Reading Test	60
Break	5
Writing and Language Test	35
Math Test – No Calculator	25
Break	5
Math Test – Calculator	45
TOTAL PSAT/NMSQT and PSAT 10	**175**

Table 4.3: Timing of the PSAT 8/9 Assessment Including Breaks

Test	Time (in minutes)
Reading Test	55
Break	5
Writing and Language Test	30
Math Test – No Calculator	20
Break	5
Math Test – Calculator	40
TOTAL PSAT 8/9	**155**

Collecting Irregularity Reports

To maintain the integrity of the assessment, it is College Board policy that all irregularities are documented, as well as any actions taken at the test center to remedy the situation. Supervisors are provided with instructions for dealing with many common irregularities. All reports of irregularities are reviewed by the College Board to determine whether the occurrence was severe enough to invalidate the test scores of the test taker(s) involved.

Complete irregularity handling procedures can be found in the Test Administration manuals.

Accommodations

The College Board believes that, in keeping with the AERA/APA/NCME Standards, all test takers should have the full opportunity to demonstrate their standing on the construct being

measured (AERA/APA/NCME, 2014). To this end, we provide testing accommodations to test takers who, due to a disability or status as an English language learner (ELL), demonstrate a need for accommodations on College Board tests. To ensure proper accommodated administrations, test takers approved for accommodations are identified at the test center via supervisor rosters and their own admission tickets. Test center staff should administer accommodated forms in accordance with the instructions provided to them pertaining to that particular accommodation. In adhering to the practice of standardization and in keeping with the Standards, all accommodated administrations are designed to be comparable to the standard administration (AERA/APA/NCME, 2014). See Section 3.5 of this manual or the test administration manuals for complete details and a listing of the types of accommodations provided.

4.2 Security

As stated previously, the College Board is dedicated to providing colleges and universities with scores that can be used for their intended purposes across all administrations and to ensuring that no test taker receives an unfair advantage on any assessment. To that end, and in keeping with the AERA/APA/NCME Standards, a number of procedures are executed to maintain test security at all times, particularly during the test administration. An overview of the most important of these procedures is provided in this section. An in-depth description of the administration procedures can be found in the test administration manuals.

All test center supervisors are trained, via written manuals and online training, to adhere to and enforce these strict security procedures.

Procedures

Several steps are taken to eliminate "opportunities for test takers to attain scores by fraudulent and deceptive means" (AERA/APA/NCME, 2014). Three important facets to the security of a test administration are preventing any test taker from having inappropriate access to the content of the assessment, confirming that the test taker who is present is indeed the person registered for the assessment, and preventing any test taker from receiving or giving assistance in completing the assessment. All of these procedures facilitate a fair testing experience for all test takers and ensure scores that are valid for intended uses across all administrations of the assessment.

The physical security of all test materials is fundamental to a fair and equitable administration. The test center supervisor is responsible for receiving the test materials, checking that they correspond with what was supposed to be shipped, and storing the materials in a locked storage area that isn't accessible to test takers or other staff. Test materials are accounted for at several points during the day of testing—when the test books and answer sheets are distributed to test takers, when they are collected from test takers, and as they are packed for return to the College Board. Supervisors are instructed to return specified test materials to the College Board immediately after the administration is complete. By maintaining the physical security of test materials, administrators make certain that no one has any access to the assessment that would provide them with an unfair advantage that might affect the scores.

For weekend administrations of the SAT, admission to test centers is carefully monitored to confirm that the individual taking the assessment is indeed the person who registered. In addition to their SAT admission ticket, test takers are instructed to bring an acceptable

photo ID, which is checked against both the admission ticket and an attendance roster previously provided to the supervisor.

For SAT School Day and the PSAT-related assessments, students who take the test at the school they attend don't produce an admission ticket or a photo ID. Students, including those who are homeschooled and who take the test at a school they don't attend regularly, are asked to display an acceptable photo ID.

Test takers aren't permitted to choose their own seats at the administration. They are assigned seating by the supervisory staff to minimize the opportunity for preplanned collaboration among friends. No unauthorized person is permitted to enter the testing room after the administration has begun.

In order to prevent test takers from receiving unfair assistance during the assessment, the materials that test takers may have on their desks during the test are limited to the test book, answer sheet, No. 2 pencils, and, for the relevant portion of the Math Test, a calculator. The only exceptions to this rule are materials approved as accommodations for test takers with disabilities or ELL needs.

Test takers are strictly prohibited from using cell phones or smartphones; audio players/recorders, tablets, laptops, notebooks, or any other personal computing devices; separate timers of any type; cameras or any other photographic equipment; any devices, including digital watches, that can be used to record, transmit, receive, or play back audio, photographic, text, or video content; protractors; compasses; rulers; dictionaries or other books; pamphlets; papers of any kind; highlighters; and colored pens or pencils. If test takers have been preapproved by the College Board for specific accommodations, they may be permitted the use of some of these materials.

Cell phones are also not allowed at the test center. We recommend that the test center staff collect them upon entry and return them after the administration.

Violation of any of these prohibitions could lead to dismissal from the testing session, cancellation of test scores, or banning from future administrations of the SAT.

Additional Test Security Procedures

Beyond the procedures undertaken at the test administration, the College Board executes several additional procedures to ensure test security. These procedures include test center audits, secure materials handling, and posttest analysis (see Section 5.2). Details of these procedures are kept confidential to maintain their efficacy as security measures.

CHAPTER 5

Interpretation and Application of Results

To ensure that scores are valid for intended uses, great care must be taken when scoring the SAT Suite of Assessments and analyzing the results. Section 5.1 describes the protocols and procedures that are followed during the operational scoring process for multiple-choice and student-produced responses, as well as the optional SAT Essay. It also discusses scale score reporting. The chapter then covers the item analysis that is performed on the operational items, including differential item functioning (DIF) and key validation. Section 5.2 describes the test security framework and procedures for the SAT Suite, including the means by which to place security holds on test taker scores that might present a plausible cause for concern. The scores then need to be reported, a process described in Section 5.3. Once students have their scores, SAT Skills Insight™ provides a set of data-driven statements intended to help students interpret their performance, a process described in Section 5.4.

5.1 Scoring Procedures

Operational Scoring Process

The AERA/APA/NCME Standards call for the establishment of scoring protocols and the documentation of quality control processes and criteria (AERA/APA/NCME, 2014). The steps include preadministration setup, scanning, production scoring for raw and scale scores, and psychometric analyses including equating. The processes are conducted through collaboration between the College Board and its subcontractors, following appropriate scoring procedures to ensure that scores are valid for intended uses.

As described in Chapter 1: Overview, the SAT Suite includes multiple-choice (MC) questions, student-produced response (SPR) questions, and an optional Essay task that is scored using a rubric.

Scantron machines that read the bubbled responses on scannable answer sheets score most MC and SPR question responses. Hand scoring, rather than machine scoring, is used for the purposes of essay scoring, quality control (QC), and, in some instances, test security. Essay scoring is performed by human raters who are trained and continually certified to score Essay responses. Hand scoring is also a way to perform QC, and is a secondary QC performed for every assessment and administration, with the primary QC checks/gates built into each of the operational systems.

Multiple-Choice and Student-Produced Response Raw Scores

Each of the MC questions includes four response choices for selection. Each of the SPR questions requires test takers to grid in up to a four-digit numerical response to an open-ended question. See Figure A-5.1 in Appendix 5: Interpretation and Application of Results for sample answer sheets for MC and SPR questions.

The MC and SPR questions are scored by reading in student responses from the gridded-in Scantron sheets, comparing the responses to preloaded keys of the correct response to

each question, and summing students' total number of correct responses in a number right scoring process. Test takers' raw responses are compared to the answer keys in the context of the entire response string and scored as 0 = wrong; 1 = right; 2 = omit; 3 = not reached; 4 = not scored. In addition, a separate set of raw item scores are created where the item responses are each scored 1 if right and 0 if otherwise.

Raw test scores and subscores are calculated by summing the number of correct responses for each of the questions contributing to that score. The Reading Test and the Writing and Language Test contain only MC questions. The Math Test has both MC and SPR questions, with each question contributing 1 point to the Math raw score.

SAT Essay Scoring

SAT Essay responses are scored using a carefully designed process. Two independent readers use the SAT Essay Scoring Rubric (see Table A-5.1 in Appendix 5: Interpretation and Application of Results) to assign a score of 1–4 in each of three dimensions: Reading, Analysis, and Writing. A score of zero is reported for an essay that is not scorable (e.g., blank, off topic, in a foreign language). If the two independent rater scores are in perfect agreement or are adjacent to each other, the scores from the two independent raters are added together to give the test taker a 2–8 score on each of the three dimensions. If the two independent raters' scores are nonadjacent (e.g., scores of 1 and 3), the Essay response is rated by a third senior-level rater. The third rater's score is then doubled to give the test taker a 2–8 score on the relevant dimension(s). These dimension scores aren't combined with each other or with scores on any other part of the SAT. The goal of reporting three separate scores is to provide students with more and better information about their performance than if they were to receive a single holistic Essay score. By evaluating students' performance in three main areas, students, parents, and educators are better able to pinpoint students' strengths and weaknesses in Reading, Analysis, and Writing.

Procedures of the Scoring Program and Process

Raters who score the SAT Essay are required to have a bachelor's degree or higher. The majority of raters selected to perform SAT Essay scoring are current teachers or those who have taught high school- or college-level courses that require writing. The raters score online remotely, under the supervision of an assigned scoring leader. Scoring leaders are selected based on their initial qualifications, as well as their highly successful performance during training, certification, and operational scoring.

During operational scoring, each rater reads a student response and evaluates the response on all three of the SAT Essay rubric dimensions—Reading, Analysis, and Writing. Raters are assigned to SAT responses randomly, according to the prompts that are the highest priority for score completion. Raters are trained to use the SAT Scoring Rubric to apply consistent and accurate scores through robust and intensive training procedures. Raters are presented with previously scored student essay exemplars for each of the score combinations possible on the SAT Essay. This enables them to become familiar with the scoring criteria in each dimension (Reading, Analysis, and Writing) at each score point and to understand how the scoring rubric's features play out in student essays at each score point. These exemplars, or anchor responses, illustrate the range of responses that fall into each score point and demonstrate how to make scoring determinations about the features that raters observe in each student essay. Each anchor essay is accompanied by a score annotation that describes the features of the essay and how those features align with the essay's score in each of the three scoring dimensions. After raters learn to assign scores according to the scoring rubric, they then practice scoring

independently, applying scores to additional prescored essays and checking their scores against score annotations. Once raters have worked through all practice materials, they must then pass a rater certification test in which the rater must assign correct scores to a prescribed percentage of prescored student essays. All raters who score operational student essays are subject to routine calibration tests in which they demonstrate their continued successful application of the rubric. Raters who do not perform adequately on these calibration tests aren't permitted to score. Raters are also subject to routine and frequent prescored validity papers that they must score accurately. Raters don't know during scoring when they are scoring a validity paper, which allows for an authentic evaluation of rater performance. Raters who don't meet acceptable measures of calibration and validity performance are released from SAT scoring.

Scale Scores Reported

Scale scores are reported at several levels to provide information that is useful for different users. Applying the appropriate raw-to-scale score conversions to the respective number right raw test scores and raw subscores creates the scale scores for test scores, cross-test scores, and subscores. The total scores and section scores are then mathematically derived from the included test scores. The graphics in Figures 5.1–5.3 present the scale score ranges (in the right-hand column) associated with each level for the SAT, the PSAT/NMSQT and PSAT 10, and the PSAT 8/9.

Initial scales were established from the winter 2014–2015 scaling study. See Section 6.1 for details on setting the initial scales for these scores and Section 6.2 for details on equating these scores across forms.

Figure 5.1: SAT scoring guide

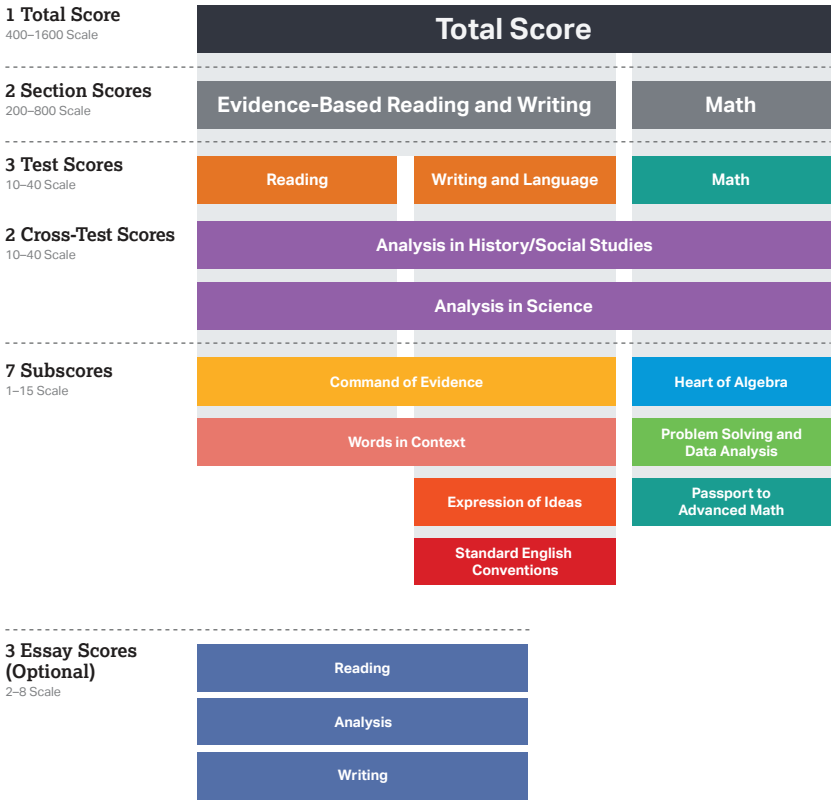


Figure 5.2: PSAT/NMSQT and PSAT 10 scoring guide

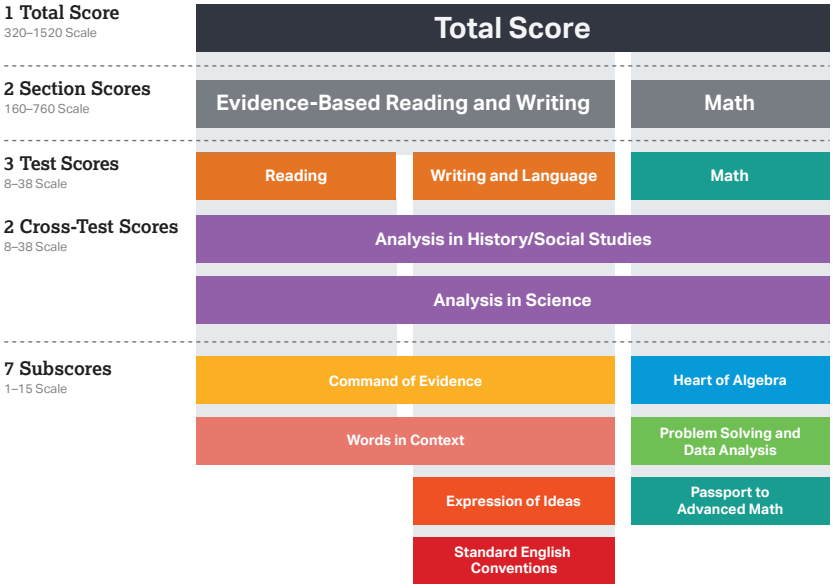
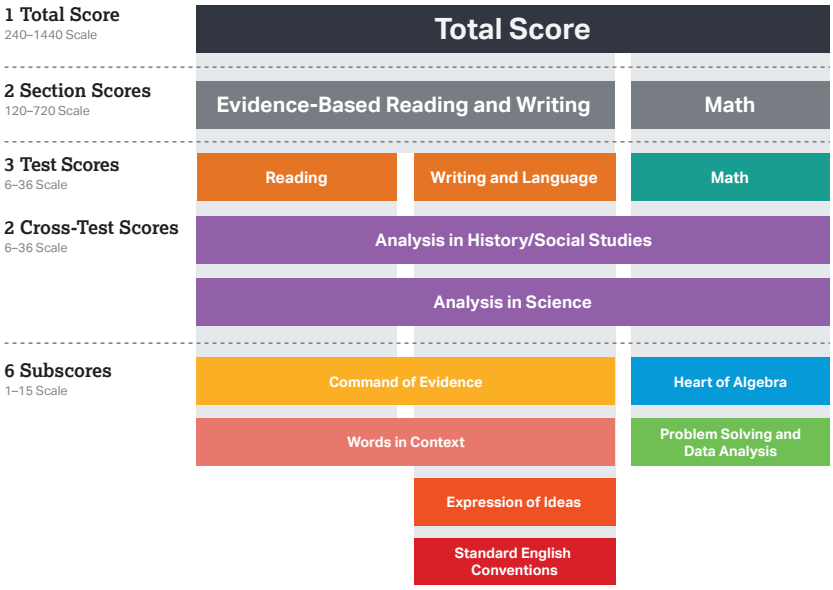


Figure 5.3: PSAT 8/9 scoring guide



Total Score

The total score is intended to be used for admission decision purposes. The total score is the sum of the two section scores: (1) Evidence-Based Reading and Writing (ERW) and (2) Math. The scores for the SAT Essay are reported separately and aren't included in the total score or section scores. The ERW and Math section scores are reported in 10-point increments on a 200–800 scale for the SAT, 160–760 scale for the PSAT/NMSQT and PSAT 10, and 120–720 scale for the PSAT 8/9. The total score is reported in 10-point increments on a 400–1600 scale for the SAT, 320–1520 scale for the PSAT/NMSQT and PSAT10, and 240–1440 scale for the PSAT 8/9. The total score isn't equated.

Section Scores

Section scores are also intended to be used for admission decision purposes. Two section scores are reported for the SAT ERW, and SAT Math. The ERW and Math section scores are reported in 10-point increments, on a 200–800 scale for SAT, 160–760 scale for PSAT/NMSQT and PSAT 10, and 120–720 scale for PSAT 8/9. The ERW section score is mathematically derived from the Writing and Language Test and Reading Test scores; it is not scaled or equated separately. The Math section score is equated (see Section 6.2 for more details).

Test Scores and SAT Essay Scores

The four test scores reported for the SAT Suite are the Reading Test, the Writing and Language Test, the Math Test, and the SAT Essay. The three non-essay test scores are equated scores and are reported on a 10–40 scale for the SAT, 8–38 scale for the PSAT/NMSQT and PSAT 10, and 6–36 scale for the PSAT 8/9. The scores are reported in 1-point increments for the Writing and Language Test and the Reading Test and in 0.5-point increments for the Math Test.

The Math Test score is derived from the Math section score, which is equated across forms. The Math Test is intended to be used for growth and accountability and is divided into two separately timed portions: a no-calculator portion and a calculator portion. Separate scores aren't reported for the no calculator/calculator portions—only a total Math Test score. The two portions are separately timed for administrative reasons, not for the reporting of scores.

The Reading Test and Writing and Language Test scores are equated across forms; they are used for growth and accountability.

The SAT Essay reports scores on three dimensions: Reading, Analysis, and Writing. These scores are reported on a 2–8 scale. A total essay score combining the three dimension scores isn't computed or reported. The essay scores are standalone and aren't combined with the Reading Test score or the Writing and Language Test score. While the three essay test scores aren't equated, tests of significance on the homogeneity of score distributions are conducted. The essay scores are intended to be used for admission decision purposes and also to help secondary students determine their strengths and weaknesses.

Cross-Test Scores

Cross-test scores are intended to be used to report on students' achievement in applying the core skills within Reading, Writing and Language, and Math to specific academic contexts. The two cross-test scores reported for the SAT Suite are Analysis in History/Social Studies and Analysis in Science. The cross-test scores comprise questions from all tests except the essay. The cross-test scores are reported in 1-point increments, on a 10–40 scale for the SAT, 8–38 scale for the PSAT/NMSQT and PSAT 10, and 6–36 scale for the PSAT 8/9. The cross-test scores are equated across forms.

Subscores

Multiple subscores are reported based on selected questions within a single test (the Writing and Language Test or the Math Test) or across multiple tests (the Reading Test and the Writing and Language Test). Two subscores are derived from questions on the Reading Test and the Writing and Language Test: Command of Evidence and Words in Context. The Writing and Language Test reports two additional subscores: Expression of Ideas and Standard English Conventions. The Math Test reports three subscores: Heart of Algebra, Problem Solving and Data Analysis, and Passport to Advanced Math.

All subscores are reported on a 1–15 scale in increments of 1. The 1–15 subscore scales are equated across forms. The subscores are intended to be used to identify strengths and weaknesses.

Item Analysis and Key Validation

While all items appearing on an operational form were pretested in some way, the functioning of the operational items are examined after each operational administration. Estimates of item difficulty and discrimination are examined for any anomalies, such as substantive changes between pretesting and operational administrations or between different forms containing the same operational items (e.g., forms containing pretest items).

See Section 3.2: Item Development for the SAT Suite Reading, Writing and Language, and Mathematics Tests for an overview of item analysis as it relates to test development. For information on DIF as it relates to fairness, see Section 2.2: Fairness Reviews of Items, Forms, and Prompts for the SAT and the PSAT-related Assessments.

Item Analyses

Listed below are item analyses conducted for the SAT Suite of Assessments. The equations can be found in Appendix 5: Interpretation and Application of Results as Exhibit A-5.1.

Item Difficulty

Item difficulties are examined using *p*-values, calculated as the proportion of test takers who correctly responded to an item out of those test takers who reached the item. With a range from 0 to 1, a lower *p*-value suggests the item is hard for test takers and a higher *p*-value suggests it is easy for test takers. Based on their sizes, the *p*-values are categorized as *easy*, *medium*, and *hard* for ease of interpretation for test takers. *P*-values are rounded to two decimal places. Tables A-5.2 to A-5.8 in Appendix 5: Interpretation and Application of Results show the frequency distributions and average *p*-values by score tier for several SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 forms.

Item Discrimination

The r-biserial correlation is a correlation between a dichotomous item score *x* (passing or not passing an individual item) and a continuous criterion score *y* (score on a subset of items). It is used as a measure of the ability of an item to discriminate between low-performing and high-performing test takers. Both the criterion score and the hypothetical ability score underlying the item score are assumed to be normally distributed. R-biserial values may range from −1 to 1. Items with higher positive values are better at distinguishing between the low- and high-performing test takers.

See Appendix 5: Interpretation and Application of Results for the formula for r-biserial correlation. Tables A-5.9 to A-5.15 in Appendix 5 show the frequency distributions and average item discrimination values by score tier for several SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 forms.

Differential Item Functioning

DIF is a statistical method that examines the performance of subgroups for possible statistical bias. Based on the formulas from Dorans and Holland (1993), the Mantel-Haenszel delta DIF (MH D-DIF) statistic is calculated for subgroups of gender and ethnicity/race. The formula for MH D-DIF is in Exhibit A-5.1 of Appendix 5: Interpretation and Application of Results.

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. During operational administration analyses, the minimum sample size requirement for the focal group is 200. Tables A-5.16 to A-5.22 in Appendix 5: Interpretation and Application of Results show the frequency distributions of the MH D-DIF statistics by content area for several SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 forms.

Key Validation

Final Score Keys and Score Reports

In cases where test takers challenge an item, staff will also review operational items and their statistics. In any situation where a problem is found with an operational item or a key, corrections are made to the key, or the item is eliminated from operational scoring. After all operational item analyses are completed and reviewed, the final score key is approved and score reports are produced. If problematic items or issues with scores are found after score reports are released, scores will be re-released after key or score changes are made.

5.2 Test Security Analyses

As mentioned in Chapter 4: Testing Requirements, posttest analysis is an important component in maintaining test security. The test security framework and procedures for the SAT include conducting broad statistical screens of score gains that don't directly result in score cancellation, but instead provide the means by which to place security holds on test taker scores, until further investigations resolve the plausible causes of concern. For the old SAT, one of these screens was a large score difference (LSD) screen, which compared a test taker's score set on an assessment and their score set on a previous administration. A score set was defined as the set of old SAT scores that a test taker received from a single administration of one assessment and was used for the score comparison analyses. For example, an old SAT score set consisted of Critical Reading, Math, and Writing scores. After each administration of the old SAT and before score reporting, each test taker's score set was compared to a previous SAT or a PSAT-related assessment score set, if available. Only one score set LSD comparison was conducted for any one test taker with multiple previous scores.

As was the case with the old SAT, score set screening is executed for all new SAT test administrations, comparing eligible students' SAT performances to previously obtained scores on either the SAT, the PSAT/NMSQT and PSAT 10 (see *Screening Eligibility*), or the old SAT. If multiple previous scores exist for a test taker, the current score set is compared to only one score set in the order of preference presented in Table 5.1.

Table 5.1: Order of Preference for Score Screening Comparisons

	Test	
Order	**Previous**	**Current**
1	SAT	SAT
2	PSAT/NMSQT and PSAT 10	SAT
3	**Old SAT**	**SAT**

If multiple previous scores exist for the same test type, only the most current score set within the test type is considered. SAT Math section scores (MSS), Evidence-Based Reading and Writing (ERW) section scores, and Total (Total) scores are compared with corresponding previously obtained SAT or PSAT/NMSQT and PSAT 10 scores, or scores from the old SAT Critical Reading (CR), Writing (W), and Math (M) section scores, as seen in Table 5.2.

Table 5.2: Score Comparisons Conducted for SAT Scores

	Previous Test Score for Comparison		
SAT Score	**SAT**	**PSAT/NMSQT and PSAT 10**	**Old SAT**
ERW	ERW	ERW	CR + W
MSS	MSS	MSS	M
Total	**Total**	**Total**	**CR + M + W**

Score Comparison Criterion

Because the new SAT scales differ from the old SAT scales, the LSD criterion couldn't be used initially. Although the PSAT/NMSQT and PSAT 10 scale is vertically aligned to the SAT scale, their score ranges differ from the SAT score ranges. An appropriate set of criteria for comparing a test taker's score set on a test and their score set on a previously administered test is required for the initial administrations of the new SAT. A static *Z*-score difference (*ZSD*) was implemented, beginning with the March 2016 new SAT administration.

The *ZSD* procedure is implemented as follows. Each selected score in the test taker's score set is standardized against the respective scale mean and standard deviation to create a *Z*-score. The standardized score from the previous administration is subtracted from the corresponding standardized score from the current administration. The difference between the current score static *Z*-score and the paired previous product static *Z*-score gives the *Z*-score difference (*ZSD*) that is investigated,

$$ZSD_{t,sc} = Z_{t,Current:sc} - Z_{t,Previous:sc} \tag{5.2.1}$$

The complete formula for calculation of the *ZSD* for any individual person and score pair, using the respective mean and standard deviation for the current and previous scores to be compared is:

$$ZSD_{t,sc} = \frac{Current:sc_t - \mu_{Current:sc}}{\sigma_{Current:sc}} - \frac{Previous:sc_t - \mu_{Previous:sc}}{\sigma_{Previous:sc}} \tag{5.2.2}$$

where *t* represents the test taker, *sc* the investigated scale score pair (three in total for each screened student) for a current SAT administration and a previous SAT, PSAT/NMSQT, PSAT 10, or old SAT administration, and μ and σ are the population scale score means and standard deviations, respectively.

For example, a student who is currently taking the SAT may have had a previous PSAT/NMSQT score set. In order to investigate the student's SAT ERW section score, both the SAT ERW and the previously obtained PSAT/NMSQT ERW section scores need to be standardized. The resulting *ZSD* for the test taker's ERW score is thus given by:

$$ZSD_{t,ERW} = Z_{t,SAT:ERW} - Z_{t,PSAT/NMSQT:ERW}$$

Screening Eligibility

For examinees to be screened using the *ZSD* methodology, they must satisfy the eligibility criteria, some of which are listed below. A full list of eligibility specifications is maintained by the College Board Business and Technical Operations Division, which oversees the security procedures for the SAT.

Eligibility Criteria:

- First-time test takers aren't screened.
- Redesigned product score comparisons are preferred (e.g., PSAT/NMSQT and PSAT 10 vs. new SAT over old SAT vs. new SAT).
- If a test taker is 13 years of age or younger, no score checks are performed.
- For students with multiple available score sets, only one comparison is carried out.
- If the comparison administrations were three or more years apart, no checks are performed.
- If the most recent scores have been canceled or have a hold placed on them, scores are compared to the next most recent available scores.
- No reverse score gains are considered; namely, if a student's performance decreased, then they aren't flagged or subject to further screening.

Each of the three score comparisons will have a *ZSD* flagging criterion. Score pair *ZSD*s must be equal to or exceed the criterion in order to result in a flag. The criteria have been selected with the goal of maintaining the historical annual flagging rate of 0.1% of test takers. Another goal of the criteria is to fairly represent the percentages of flags across subjects (ERW and MSS).

5.3 Reporting

Score reporting

In keeping with the AERA/APA/NCME Standards, SAT Suite score reports have been developed at the student and institutional levels to provide their intended audiences with appropriate interpretations of the reports and guidelines outlining the appropriate use of test results.

Student Score Reports

Online and paper score reports are available to students taking the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9.

Online Score Reports

Online score reports are available to students registering for the SAT online or who registered by mail and set up a College Board account. PSAT-related online score reports are available to students who set up a College Board account. Students must be 13 or older to set up a College Board account.

To explore the live online student demo account, go to www.sat.org/scoresdemo.

Paper Score Reports

SAT paper score reports are available to students who register online or by mail. Students registering online must request to receive a paper score report by mail in addition to their online score report. The students registering by mail and who don't have active College Board accounts will receive paper score reports, provided the College Board has an address on file. For PSAT-related assessments, paper score reports are sent to the student's school for distribution.

A sample PDF of the paper score report is downloadable from the demo account referred to in the online score reports section.

Features of Student Score Reports

Student score reports display scores on each of the scale scores of the SAT or PSAT-related assessments, including the total, section, test, cross-test, and subscore, providing details to students on how they performed in specific areas (displayed in online and paper score reports).

Score ranges are displayed online with every score, indicating how students would perform if they were to take the same test repeatedly with no new learning. The score range is computed as the student's score +/− the standard error of measurement (SEM). See Section 6.4: Reliability for more information on SEM computation. Paper reports have generalized SEM statements for each score tier, for example, "Subscores: Your Score +/− 2 points."

Percentile ranks allow students to compare their scores to those earned by other students. Two percentile ranks are displayed online in student score reports:

- The Nationally Representative Sample Percentile shows how your score compares to the scores of all U.S. students in your grade, including those who don't typically take the test.
- The SAT User Percentile shows how your score compares to the scores of students who typically take the test.

Due to space constraints, paper reports only report a subset of the percentiles. For more information on the development of the percentile ranks, see Section 6.3: Normative Information.

The SAT College and Career Readiness Benchmarks are displayed at the section level (Evidence-Based Reading and Writing, Math) and are based on actual student success in entry-level college courses. The SAT benchmark scores represent a 75% likelihood of a student achieving at least a C in a first-semester, credit-bearing college course in a related subject. The Evidence-Based Reading and Writing benchmark is the score associated with a 75% chance of earning at least a C in first-semester, credit-bearing, college-level courses in history, literature, social science, or writing (displayed in online and paper score reports). The Math benchmark is the SAT Math section score associated with a 75% chance

of earning at least a C in first-semester, credit-bearing, college-level courses in algebra, statistics, precalculus, or calculus.

The grade-level benchmarks for the PSAT-related assessments are based on expected student growth toward the SAT benchmarks at each grade. Whereas SAT benchmarks indicate likelihood of success in college, grade-level benchmarks indicate whether a student is on track for college and career readiness for their grade.

For more information on the SAT benchmark, see Section 7.5: Measuring and Monitoring College Readiness with the SAT. More information about the college and career readiness benchmarks can also be found at **https://collegereadiness.collegeboard.org/pdf/educator-benchmark-brief.pdf**.

Red, yellow, and green score indicators provide insightful feedback for reported scores, indicating specific areas of strength and weakness. To provide students and educators with tools and guidance on where to focus studies to meet the benchmarks, the College Board has developed a color-coded system for reporting section scores, test scores, and subscores. When students view their score report, section scores will be accompanied by graphics indicating their score and its relation to the corresponding benchmark. Scores that meet or exceed the benchmark in the Evidence-Based Reading and Writing section or the Math section will be marked in a green area on a score graphic. Scores that are below the benchmark but within one year's growth of meeting the benchmark are marked on the yellow portion of the graph. Scores that are below the benchmark and are more than one year's growth from meeting the benchmark are indicated on the red portion of the graph. Color coding for test scores and subscores provides indicators to inform students how their performance on the test scores and subscores compares to students who are on track for college readiness. The red, yellow, and green ranges for test scores and subscores are based on the average performance of all test takers who met the corresponding section benchmarks for their grade level compared to those who didn't. The red color indicator tells students that they need to strengthen their skills, the yellow tells them that they are approaching the benchmark, and the green that they have performed as well as, or better than, students who met or exceeded the benchmark (displayed in online score reports).

SAT Essay scores (only available on the SAT) are displayed to show performance on the Reading, Analysis, and Writing sections of the Essay, which are scored on a scale from 2 to 8 (displayed in online and paper score reports).

Skills Insight is designed to help students acquire a better understanding of how scores relate to specific academic skills. It offers descriptions of performance and insight into skills measured at each score band. It also provides actionable suggestions for improving skills that help students gain additional practice (displayed in online score reports only). For more information on Skills Insight, see Section 5.4: SAT Skills Insight.

The College Board has partnered with Khan Academy to provide free personalized practice recommendations online. Official SAT Practice on Khan Academy provides students with official practice tests and study/test-taking tips, diagnostic quizzes to identify areas for further practice, practice questions, video lessons, hints, and constant feedback and progress.

K–12 Institutional Score Reports

K–12 Reporting Portal

Scores for the SAT, SAT Subject Tests, PSAT/NMSQT, PSAT 10, and PSAT 8/9 are released through the new integrated score reporting portal (https://collegereadiness.collegeboard.org/educators/k-12/reporting-portal-help), which supports effective decision making with standard reports, interactive data analysis tools, and secure file downloading.

Report Types

Scores by Institution provides aggregate performance data by institution, as well as student-level performance at each school for every score.

Scores by Demographics provides aggregate performance data by demographic group for every score. It also displays aggregate performance data for the entire institution and allows for the reporting of one or two demographic groups at a time.

Benchmarks by Institution measures aggregate as well as student-level performance against College and Career Readiness Benchmarks for the Evidence-Based Reading and Writing section and the Math section.

Benchmarks by Demographics provides aggregate performance against benchmarks for the Evidence-Based Reading and Writing section and the Math section by demographic group. It also displays aggregate performance data for the entire institution and allows for the reporting of one or two demographic groups at a time.

Instructional Planning highlights instructional strengths and weaknesses and can be used to help focus improvement efforts. It measures aggregate performance data against the benchmarks for every score except the total score and cross-test scores.

Question Analysis provides aggregate and student-level performance data for every question on the assessment. It can be used to learn how students answered each question, which mistakes were the most common, how well each skill has been mastered, and which state standards have been met (not yet available for all states).

Roster Summary provides aggregate and student-level test administration information, including registration, fee waiver usage, absenteeism, and scores for each student registering for and/or completing the test.

Electronic Score Reports (or Data Files) are available as of the new SAT and PSAT-related administrations. Electronic score data files are available to educators in the online score reporting portal's Download Center and are available in TXT and CSV file formats for integration into existing K–12 reporting systems (e.g., Naviance).

Designated Institution Score Reports

Students can opt to send their SAT scores to colleges, scholarships, and other designated institutions via official score reports. Colleges and scholarship administrators use SAT scores for admission applications and other opportunities.

Colleges can choose to receive paper or electronic score reports (data files). Electronic score data files are available to higher education institutions in the Higher Education Reporting portal (https://collegereadiness.collegeboard.org/educators/higher-ed/reporting-portal-help).

The Higher Education Reporting portal provides additional reports and features to higher education institutions:

- SAT Trend Dashboard that provides year-over-year SAT score send trends.
- View student SAT essays online or generate and download SAT essay batch files to load into student information systems.
- Enrollment Planning Service (EPS) subscribers can also view Executive Summary reports.

5.4 SAT Skills Insight

Skills Insight provides a set of data-driven statements intended to help students interpret their performance on the Reading, Writing and Language, and Math Tests of the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9. The statements are organized by test score band: 6–14, 15–19, 20–24, 25–29, 30–34, and 35–40. Within each score band, the Skills Insight statements describe what a student scoring within that band is likely to know and be able to do in relation to the academic skills measured on the tests. The goal of Skills Insight is to help students, teachers, administrators, and others understand what a test score means and, for scores below the highest range, how performance could be improved.

Determining Score Ranges

Before Skills Insight statements could be developed, it had to be determined what score ranges (i.e., score bands) would be the most effective in providing narrative information to students about the skills they most likely had mastered. These score ranges were determined by first reviewing the test score scale underlying the SAT Suite of Assessments to ensure adequate coverage of the scale for the full suite. Having too many score bands wouldn't allow us to provide skills that were meaningfully different between score groups, and, therefore, wouldn't provide useful information to the students. Having too few score bands wouldn't allow us to pinpoint the strengths and weaknesses of the students represented in the different score bands. The distribution of scores across the scale covering the full SAT Suite was also considered when determining the number of score bands and the width of each band. Given that the preliminary set of Skills Insight statements was developed before operational data on the tests were available, data from forms administered in the initial research studies conducted prior to the first operational administration were used.

After analyzing all the data and reviewing different possible score band numbers and widths, College Board staff concluded that the test score ranges 6–14, 15–19, 20–24, 25–29, 30–34, and 35–40 would best distinguish students' skills so as to be helpful to them as they plan their practice in the skills that are essential for being college and career ready.

Initial Development of the Skills Insight Statements and Suggestions for Improvement

The preliminary Skills Insight statements were developed in summer 2015 from the score data of large representative samples of students who took one of the assessments as administered in one of several research studies conducted prior to the first operational administration of the forms. Multiple forms of each assessment were reviewed—six SAT test forms, three PSAT/NMSQT and PSAT 10 test forms, and two PSAT 8/9 test forms. For a given subject—Reading, for example—each question was placed into one of six groups corresponding to the six score bands listed previously. The score range to which a question

was assigned was the lowest score band in which the question was answered correctly by at least 75% of students scoring in that band.

College Board content area staff drafted Skills Insight statements by considering which academic skills students would need to have mastered in order to be likely to answer successfully the items assigned to each of the score ranges 6–14, 15–19, 20–24, 25–29, 30–34, and 35–40. Statements were consolidated to assure that Skills Insight reflected the continuum of skills needed to be successful throughout the score range of the SAT Suite. College Board curriculum and instruction staff developed suggestions for improvement at each score band except the highest to describe steps students could take to acquire the additional academic skills needed in order to score in the next higher score band.

Validation of the Skills Insight Statements and Suggestions for Improvement

To validate the preliminary Skills Insight statements, data from operational administrations of 16 additional SAT test forms and three additional PSAT/NMSQT and PSAT 10 forms were included in a further review by College Board staff in fall 2016. Following the procedure described earlier in this chapter for assigning questions from the additional forms to the test score ranges, the placement of questions was found to be consistent with the Skills Insight statements in each score range. Table A-5.23 in Appendix 5: Interpretation and Application of Results shows the total number of forms and questions reviewed for each assessment within the SAT Suite of Assessments included in the development of Skills Insight.

Independent Review of Skills Insight Statements and Suggestions for Improvement

To independently validate the Skills Insight statements, the College Board conducted an external review in late 2016 by recruiting current subject-matter experts knowledgeable in the instruction of reading, writing and language, and math at the middle school, secondary, and postsecondary levels. The selection process yielded a diverse group of experts in terms of gender (31% female, 69% male), ethnicity/race (23% Hispanic, 15% African American, 8% Asian American, 54% White), and geographic location (representing all geographical regions within the United States). The external reviewers often had expertise at more than one level: 61% had experience at the postsecondary level, 54% at the secondary level, and 15% had also taught middle school.

The subject-matter experts were asked to consider whether the Skills Insight statements were clear, relevant to the curriculum, and aligned in a continuum from lower to higher levels of achievement when progressing from lower to higher score bands. The reviewers also examined whether the Skills Insight statements reflected the skills and knowledge required to score in a particular range, consistent with the test questions that had been assigned to each score band (10–20 exemplar questions for each score band). Another task for the subject-matter experts was to review the Skills Insight suggestions for improvement for clarity, appropriateness, and whether the suggestions (if followed) would help students improve their academic skills. Finally, they were asked to provide revisions to any of the statements and suggestions for improvement they reviewed that would improve the use of Skills Insight as an interpretative tool.

Expert judgments from the external review were tabulated as follows. For each content area, the College Board staff calculated the percentage of reviewers in agreement that the placement of Skills Insight statements within each score band reflects appropriate

academic skills, the percentage of reviewers in agreement that the questions placed in the score band are consistent with the Skills Insight statements, and the percentage of reviewers in agreement that the suggestions for improvement at each score band reflect sound ideas for the development of academic skills needed to achieve a score in the next higher score band. Table A-5.24 in Appendix 5: Interpretation and Application of Results summarizes the judgments of the subject-matter experts. The external reviewers found the Skills Insight statements to be well aligned with expectations for academic skills in the score bands, the questions placed into the score bands to be very consistent with the Skills Insight statements within those bands, and the suggestions for improvement to be generally accurate and helpful guidance to students in their preparation for retesting.

External reviewer comments about Skills Insight in general and specific revisions to the statements and suggestions for improvement were useful and considered by College Board content area staff in further refinement of SAT Skills Insight.

CHAPTER 6

Psychometrics

Having discussed the gathering and analysis of scores in the previous chapter, we must now use those scores to establish numerical systems that convey test performance. Within this chapter, Section 6.1 describes how we established a measure that would allow us to better present appropriate interpretations of SAT performance, known as a scale. Steps must then be taken to maintain the scales that we established, a process known as equating. Section 6.2 describes our equating procedure, which accounts for differences in difficulty and is used to make the scale scores interchangeable across multiple test forms and test dates. We then turn our attention to where typical and national groups fall on the scales we created during scaling. Section 6.3 describes the creation of the normative information used to support appropriate interpretations of the common score scales among College Board's constituents. The next step in the process is to determine the precision of the measures we are using. To this end, Section 6.4 describes the reliability procedures used to measure the consistency in test takers' observed scores across instances of the test procedure. Besides the aforementioned sections that address analyses of test scores, there are additional psychometric analyses that have particular applications to overall processes of item development, test form assembly, and test score evaluation, which are discussed in Section 6.5.

Analyses described in this chapter and in Section 7.3 were based on data sets, studies, and test administrations that took place from 2014 to 2016. A large national sample of high schools and students took initial test forms of the redesigned SAT and PSAT-related assessments in 2014, prior to the launch of the redesigned SAT Suite of Assessments, and provided the data used to establish the SAT scales (Section 6.1, the 2014 Scaling Study), equate the initial SAT and PSAT-related forms (Section 6.2), estimate nationally representative norms (Section 6.3), and establish the vertical scales for the PSAT-related forms (Section 7.3). A more recent data set of examinees who took the old and redesigned SAT tests were used to estimate SAT user norms (Section 6.3) and also produce the SAT concordances (Section 7.3). The user norms for the PSAT-related assessments reported in Section 6.3 were estimated using administration data from the 2015–2016 redesigned PSAT-related assessments. Finally, reliabilities, and item and score statistics described in Sections 6.4 and 6.5 were estimated from actual administrations of the redesigned SAT.

6.1 Scaling

As stated earlier in this chapter, the purpose of scaling is to establish numerical systems that convey test performance. The best scales are the ones that support intended interpretations of test performance, which for SAT involves the scale score systems summarized in Section 1.3 and Section 5.1. The most significant parts of this scaling work began in December 2014, when the College Board conducted a large study with a group of nationally representative high school students. Using data from the 2014 Scaling Study, the College Board established 12 separate scores of the new SAT base form (Base Form A in Section 6.2). These scores

included the Math section score, the Reading Test score, the Writing and Language Test score, the Analysis in Science and Analysis in History/Social Studies cross-test scores, and seven subscores. This section describes the goals for establishing the new SAT scales, the data collection for the scaling study, the scaling process, and the results (AERA/APA/NCME, 2014). Additional work for evaluating the scales is also discussed. The scores from the PSAT-related assessments are vertically scaled, which is discussed in-depth in Section 7.3.

Goals for the Scales

The scale scores were established as conversions of the number correct scores for 12 scores of the new test. This process was based on goals consistent with how the scores were intended to be established, as described in Sections 1.3 and 5.1.

Math and Evidence-Based Reading and Writing (ERW) section scores with:

- Ranges of 200–800
- Means of 500 for a college-bound group weighted to reflect the old SAT cohorts
- Distributions that are similar with respect to standard deviations (about 100) and skewness
- Conditional standard errors of measurement (CSEMs)[1] that are approximately constant and similar along the entire score range
- All correct, maximum possible raw scores convert to a highest obtainable scale score of 800
- None correct, minimum possible raw scores convert to a lowest obtainable scale score of 200
- Minimized gaps and many-to-one conversions in the rounded raw-to-scale score conversion tables

Math, Reading, Writing and Language Test and Science and History/Social Studies Cross-Test Scale Scores with:

- Ranges of 10–40
- Means of 25 for a college-bound group weighted to reflect the old SAT cohorts
- Distributions that are similar with respect to standard deviations (about 5) and skewness
- CSEMs that are approximately constant and similar along the entire score range
- All correct, maximum possible raw scores convert to a highest obtainable scale score of 40
- None correct, minimum possible raw scores convert to a lowest obtainable scale score of 10
- Minimized gaps and many-to-one conversions in the rounded raw-to-scale score conversion tables

[1] Standard errors of measurement reflect imprecision in test scores due to the particular sample of items on the test form. This type of error differs from the standard errors discussed in Equating (Section 6.2) and Normative Information (Section 6.3) in that those were standard errors due to sampling that reflect samples of examinees.

Subscores with:

- Ranges of 1–15
- Means of 8 for a college-bound group weighted to reflect the old SAT cohorts
- CSEMs that are approximately constant along the entire score range
- Distributions that are similar with respect to standard deviations
- All correct, maximum possible raw scores convert to a highest obtainable scale score of 15
- None correct, minimum possible raw scores convert to a lowest obtainable scale score of 1
- Minimized gaps and many-to-one conversions in the rounded raw-to-scale score conversion tables

The scaling goals for the sections, tests/cross-tests, and subscores are intended to support appropriate interpretations of SAT test performance. Scale score ranges and minimum and maximum possible scores reflect the primary expectation of how the new SAT scales are presented to test takers (Sections 1.3 and 5.1). The scales are relatable across the sections, tests, cross-tests, and within subscores, and are not easily confused with number-correct scores or the scales of other testing programs. Minimizing gaps and many-to-one conversions in the rounded raw-to-scale score conversion tables encourage sound score interpretations and differentiations among test takers. Approximately equal measurement precision in terms of stabilized *CSEMs* was also a high-priority scaling goal because it supports scale score interpretations with respect to a single standard error of measurement value rather than multiple *CSEMs*.

Sample Design and Data Collection for the Scaling Study

Several new SAT forms were administered in the 2014 Scaling Study. After evaluation of their statistical properties, one form was identified as the base form for the SAT. Data were collected using the SAT base form (referred to as Form A in Section 6.2) for the purpose of scaling, and a self-reported survey was administered with the SAT base form. There were 6,024 nationally recruited 11th- and 12th-grade test takers who took the SAT base form in the scaling study. Nationally recruited test takers were obtained from a list of high schools selected to achieve a nationally representative sample of high schools in terms of variables such as College Board region, ethnic distribution, percentage of students receiving free and reduced price lunches, and school type (private/public). From that group of test takers in the sampled high schools, we composed a sample that represents a typical SAT cohort group by identifying motivated test takers who were similar to the old SAT cohorts in terms of several background variables. This process is described in detail in the following paragraphs.

The scaling study sample was evaluated based on test takers' responses to the test items and the survey. Motivated and unmotivated test takers were identified based on their percentages of completed test items and also on their responses to a survey question about their effort given on the SAT tests. Other survey questions were used to identify a college-bound group that is composed of 11th and 12th graders. Eight samples were initially examined by considering completion rate, survey questions on test takers' motivation, grade level, and educational plans beyond high school. The final sample of 4,346 test takers was selected because it was the most desirable in terms of sample size, statistics, cohort representativeness, and scaling feasibility.

Lastly, we examined the weighted sample of 11th- and 12th-grade test takers after screening for motivation. To compose the weighted sample, the weighting method described in Haberman (1984, 2015) was used. The purpose of the weighting was to create a representative sample of the SAT cohort group by approximating an average of the 2011–2014 SAT cohorts with respect to subgroup percentages on several background variables, including percentages of 11th and 12th graders, ethnicity subgroups, gender, test takers' college plans, College Board region, mother's and father's education, first and best language, grade point average (GPA), and honors/AP coursework in math and English.

Table A-6.1 in Appendix 6: Psychometrics shows the average subgroup percentages of these background variables for the old SAT Cohort sample. The percentages of subgroups in the unweighted scaling sample were quite different from the percentages in the weighted scaling sample. For example, the percentage of 11th graders in the unweighted scaling sample was 67%, where it was 33% in the weighted sample. In fact, the percentages of subgroups in the weighted scaling sample were almost identical to the ones in the SAT cohort. Thus, the goal of weighting to approximate the distributions of SAT cohort characteristics was successfully achieved.

Scaling Procedure

The scales for the Math section score, the Reading Test score, the Writing and Language Test score, the two cross-test scores, and seven subscores were established using methods that stabilize the CSEMs across the raw scores. The Evidence-Based Reading and Writing (ERW) section scores were mathematically derived from the rounded Reading Test (R) and the Writing and Language Test (WL) scale scores with mathematical expressions corresponding to the descriptions in Section 5.1,

$$ERW = R \cdot 10 + WL \cdot 10 \tag{6.1.1}$$

The Math Test (MTS) scale scores were derived from the rounded Math section scale scores (MSS),

$$MTS = \frac{MSS}{20} \tag{6.1.2}$$

The total scale scores were derived from the ERW and Math section scores,

$$\text{Total} = ERW + MSS \tag{6.1.3}$$

The two methods considered for the scaling procedure were the arcsine transformation and a cubic transformation obtained from numerically minimizing *CSEM*s estimated from the compound binomial model. For the arcsine transformation method, the raw scores are transformed using the following equation:

$$g(y) = 0.5\left\{\sin^{-1}\left[\left(\frac{y}{K+1}\right)^{\frac{1}{2}}\right] + \sin^{-1}\left[\left(\frac{y+1}{K+1}\right)^{\frac{1}{2}}\right]\right\} \tag{6.1.4}$$

where

y is the raw score,

K is the number of items on the test, and

$\sin^{-1}$ is the arcsine function (Kolen & Brennan, 2014, p. 405).

To obtain the desired mean and average Conditional Standard Error of Measurement (*CSEM*), the scale scores, *sc*[*g*(*y*)], can be found by linearly transforming the arcsine transformed scores as follows:

$$sc\left[g\left(y\right)\right] = \frac{sem_{sc}}{\widehat{sem_g}} g\left(y\right) + \left\{\mu_{sc\left[g(y)\right]} - \frac{sem_{sc}}{\widehat{sem_g}} \overline{g\left(y\right)}\right\} \tag{6.1.5}$$

where

g(*y*) is the arcsine transformed score,

$\widehat{sem_g}$ is the estimated average standard error of measurement (*SEM*) of the arcsine transformed scores,

sem_{sc} is the desired average *SEM* of the scale scores,

$\mu_{sc[g(y)]}$ is the desired scale score mean, and

$\overline{g(y)}$ is the estimated mean of the arcsine transformed scores (Kolen & Brennan, 2014, p. 407).

For the SAT scaling, the compound binomial model was used to estimate $\widehat{sem_g}$, the *SEM* of the arcsine transformed scores. The details of the arcsine transformation stabilizing scale score *CSEMs* can be found in Kolen and Brennan (2014).

As an alternative method to the arcsine transformation method, a numerical approach to stabilizing *CSEMs* in the raw-to-scale score transformations was also considered in SAT scaling (Moses & Kim, 2017). In the cubic transformation method, a raw-to-scale score transformation is defined as a cubic polynomial for producing scale score distributions:

$$sc\left(y\right)_{Cubic} = \delta_0 + \delta_1 y + \delta_2 y^2 + \delta_3 y^3 \tag{6.1.6}$$

where *y* is the raw score and the δ's are the polynomial coefficients (Moses & Golub-Smith, 2011). The values of the δ's are numerically solved not only to produce scale scores with a desired mean and standard deviation but also to minimize a function of scale score *CSEM* instability defined as

$$\sum_{y=1} \left| CSEM_{sc\left(y\right)_{Cubic}} - CSEM_{sc\left(y-1\right)_{Cubic}} \right| \tag{6.1.7}$$

For the technical details of the cubic transformation, refer to Moses and Golub-Smith (2011) and Moses and Kim (2017).

Two scaling methods were implemented to produce section, test, and cross-test scales with desired means and standard deviations and minimized gaps and many-to-one conversions. To achieve the goals of the SAT scaling, the iterative process of SAT scaling was conducted by

examining multiple levels of *CSEMs* for the arcsine transformation method and different standard deviations for the cubic transformation method. The scaling methods ultimately selected were those that produced the most similar means and standard deviations of the Math and ERW section scores, the three test and two cross-test scores in the weighted data. Linear interpolation adjustments were applied to the lowest obtainable scale score and the highest obtainable scale score to produce more desirable highest and lowest scale score conversions, and also to prevent the unrounded scale scores from being extremely outside of the established ranges.

The scales for the seven SAT subscores were established by following similar iterative scaling procedures to the ones employed for test and cross-test scores using the two scaling methods—arcsine transformation and cubic transformation methods. The scaling methods that produced scales with the fewest gaps and many-to-one conversions in the 1–15 ranges were selected as the final methods. Linear interpolation adjustments were applied to the highest and lowest scale scores to produce more desirable highest and lowest scale score conversions, and also to prevent the unrounded scale scores from being extremely outside of the established ranges.

Results

After conducting the iterative process of SAT scaling described earlier in this chapter, the raw-to-rounded scale score conversions for the 12 SAT scale scores were developed. The arcsine transformation method was used to set the scales for the Reading Test and the Analysis in Science Test, while the cubic transformation was used for the Math section, the Writing and Language Test, and the Analysis in History/Social Studies test scores. For all seven subscores, the arcsine transformation method was used and the following desired *CSEM* values were selected for each subscore:

- Command of Evidence (COE): 1.3
- Words in Context (WIC): 1.7
- Standard English Conventions (SEC): 1.2
- Expression of Ideas (EOI): 1.5
- Heart of Algebra (HOA): 1.4
- Passport to Advanced Math (PAM): 1.6
- Problem Solving and Data Analysis (PSD): 1.5

Table A-6.2 in Appendix 6: Psychometrics shows the summary statistics of SAT rounded scale scores based on the weighted sample with the weighted sample size scaled to sum to N = 4,346. The scale score means for section scores, test scores, cross-test scores, and subscores were almost identical to the target mean scores of 500 for section, 25 for test and cross-test, and 8 for subscores. In addition, all scale scores appeared to have similar average *CSEMs* and similar standard deviations across all scale scores of any given type. The *CSEMs* for the scales that were directly established from SAT scaling—Math section score, Reading Test score, Writing and Language Test score, two cross-test scores, and subscores—were computed based on the method described in Kolen, Hanson, and Brennan (1992).

The plots of scale score *CSEMs* for Math, Reading, Writing and Language, Science, and Analysis in History/Social Studies are presented in Figures A-6.1 through A-6.5 in Appendix 6: Psychometrics. As shown in these figures, the *CSEMs* for all test and

cross-test scores were approximately constant across all scores. Figure A-6.6 in the same appendix shows the plots of linearly adjusted and unrounded scale score *CSEMs* for the seven subscores, while Figure A-6.7 shows the plots of adjusted and rounded scale score CSEMs. The unrounded scale score *CSEMs* were approximately constant across all seven subscores. On the other hand, the *CSEMs* of adjusted and rounded scale scores for some subscores were slightly inconsistent for certain ranges of scale scores. This was mainly due to the adjustment for the highest and lowest scale scores, truncation and rounding of the unrounded scale scores, and presence of a smaller number of items compared to test and cross-test scores.

Estimation equations for scale score reliabilities were developed based on *CSEMs* for the established scales, as reported in this section and also in Section 6.4:

$$\text{Reliability}_{sc} = 1 - \frac{MS\,(CSEM)_{SC}}{SD^2_{SC}} \tag{6.1.8}$$

where SD^2_{SC} is the estimated variance of scale scores. Cronbach alpha was reported for raw score reliability. The mean squared *CSEM*, *MS*(*CSEM*), was obtained as the weighted average of the squared *CSEMs* for the scales directly established. Thus, the *MS*(*CSEM*) can be written as

$$MS\,(CSEM)_{sc} \approx \int CSEM^2_{sc(\tau)} \text{Prob}(\tau)\, d\tau \tag{6.1.9}$$

where $CSEM^2_{sc(\tau)}$ is the squared scale score *CSEM* at τ, and the average of these is obtained over the probability distribution of τ, *Prob*(τ).

For the scores that were mathematically derived, including Math Test (Equation 6.1.2), ERW (Equation 6.1.1), and total scores (Equation 6.1.3), the following equations were used to compute the root mean squared *CSEM*, *RMS*(*CSEM*):

$$RMS(CSEM)_{MTS} = \sqrt{\frac{MS(CSEM)_{MSS}}{20^2}} \tag{6.1.10}$$

$$RMS(CSEM)_{ERW} = \sqrt{MS\left(CSEM\right)_R \cdot 10^2 + MS\left(CSEM\right)_{WL} \cdot 10^2} \tag{6.1.11}$$

$$RMS(CSEM)_{Total} = \sqrt{MS\left(CSEM\right)_{ERW} + MS\left(CSEM\right)_{MSS}} \tag{6.1.12}$$

Evaluations of the SAT Scales

Since their initial establishment, the SAT scales have been evaluated in several ways and have been studied in terms of how well they support equating of the alternate forms (see Section 6.2). The scales have also been evaluated in terms of meeting the scaling goals with respect to *CSEMs*, minimum and maximum possible scores, etc.

The AERA/APA/NCME Standards call for warning test users of the limitations and potential misinterpretations of the reporting scales (AERA/APA/NCME, 2014). A particularly important limitation and potential misinterpretation of the SAT scales is that the scales weren't

established based on the performance of actual SAT test takers. Instead, the scales were established by indirectly approximating the test performance of a target population of interest, the historical SAT cohort. This approximation was based on voluntarily participating, nationally recruited high school students, where the actual motivation of test takers was approximated through data screenings and other incentives, and where demographic characteristics were approximated through weighting. The screenings and weightings were selected from several plausible options. Because of the limitations of this approach, goals for section score means of 500 and test and cross-test score means of 25 may not be met in actual SAT administration data, and strong interpretations of test scores with respect to these targets may be inaccurate.

Sound psychometric practice for testing programs calls for periodic checks of their reporting scales for stability (AERA/APA/NCME, 2014). The scales established for the new SAT should be continually evaluated for indications that revisions are warranted.

6.2 Equating

It is common practice to develop multiple test forms in large-scale assessment programs in order to support numerous test administrations over several years. This collection of test forms is purposefully developed to be parallel by building the forms according to the same test specifications and making them comparable in terms of difficulty. However, equal form difficulty can rarely be achieved in practice, due to various external constraints. As a result, the difficulty level of these forms varies to some degree. Equating is a statistical procedure that accounts for differences in difficulty and is used to make scores interchangeable across multiple test forms (AERA/APA/NCME, 2014). Equating in this case assumes that a score scale is already available.

Equatings of the first redesigned SAT forms were conducted by the College Board's psychometric staff in June 2015, immediately after the base scales were constructed using test taker data for a base form and for other forms collected in the Scaling Study conducted for setting the SAT scales (see Section 6.1). This section provides information about all aspects of the psychometric work, including the equating design, the methods, the characteristics of the equating sample, and scale score information; the section also provides a set of sample SAT conversion tables. A brief overview of the College Board's internal equating system concludes the section.

Equating Design

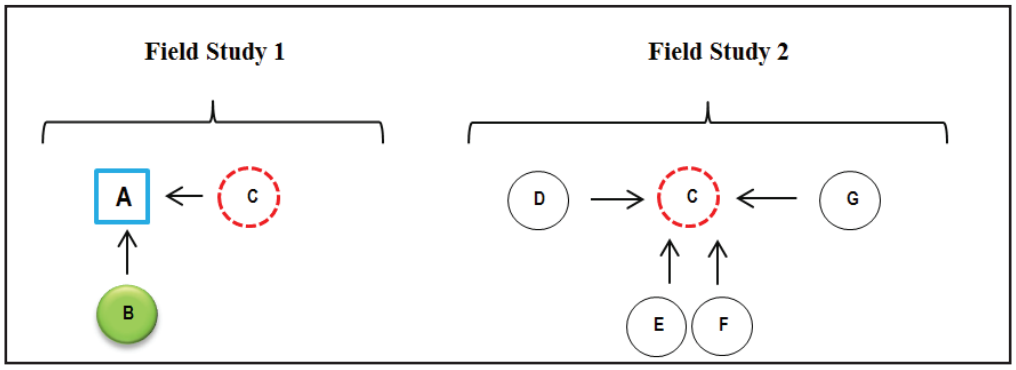
The College Board's psychometric staff conducted three major standalone field studies in 2014 and 2015. The goals of these field studies were to (1) develop new score scales, (2) equate the first set of SAT forms to be available in future administrations, and (3) build concordance tables to link the old SAT to the new SAT.

Equating was conducted through a twofold study that included seven test forms. As shown in Figure 6.1, two new forms were built into Field Study 1 (also cited as Scaling study), where equating used a nationally representative sample, primarily collected to establish the base scales. (Details about scaling analysis are in Section 6.1: Scaling.) New forms were also placed in Field Study 2 (also cited as Equating study), where data were collected specifically for equating.

A random groups (RG) data collection design was used to equate the new forms; thus, all forms were spiraled *within* each study and then linked through an anchor form. In Field study 1, two new

forms (Form B and Form C) were spiraled in data collection with the base form (Form A), which originally was used to set the new SAT scales. Field study 2 included four new forms (Forms D, E, F, and G). Form C, which was equated in Field Study 1, served as the anchor form, providing a link between the forms used in these two field test studies. The rest of the section provides discussion about equating of a new SAT form (Form B) through a base form (Form A) in detail.

Figure 6.1: Equating design



Key: Square = scaling form, Circle = new form (to be equated), Dashed circle = anchor form. All colors are arbitrary.

As a result of SAT equating, 15 scores become available for reporting (described in detail in Section 5.1). Twelve of these scores are equated and three of them are derived. Equating is performed for each of the following scores:

1. Math (section score)
2. Reading (test score)
3. Writing and Language (test score)
4. Analysis in Science (cross-test score)
5. Analysis in History/Social Studies (cross-test score)
6. Command of Evidence (subscore)
7. Relevant Words in Context (subscore)
8. Expression of Ideas (subscore)
9. Standard English Conventions (subscore)
10. Heart of Algebra (subscore)
11. Problem Solving and Data Analysis (subscore)
12. Passport to Advanced Math (subscore)

Three scores that aren't equated but rather derived from the equated and scaled scores are:

(1) Math Test (MTS) score is derived from equated Math section scale score (MSS):

$$MTS = \frac{MSS}{20}$$

(2) Evidence-Based Reading and Writing (ERW) section score is derived from rounded Reading (R) and Writing and Language Test (WL) scale scores:

$$ERW = R \cdot 10 + WL \cdot 10$$

(3) Total score is derived from Evidence-Based Reading and Writing and Math section scores:

$$\text{Total} = ERW + MSS$$

In the following text, we will discuss the equating of a redesigned SAT form (Form B) in detail—equating information regarding other forms is not included here.

Equating Sample

Building randomly equivalent groups can be operationally complex, as form spiraling must result in test taker groups that are equivalent in ability in order to provide a common link between the base form and the new forms. The key element of the RG design is the precise form assignment (spiraling) within each participating school. Therefore, to make sure that form assignment was executed as expected, distribution patterns and related statistical properties were analyzed for each set of forms that was spiraled within each school. If any school is identified as an anomaly regarding form assignment, the school will be excluded from the equating sample.

Furthermore, to address any potential problem with decreased motivation level of participating test takers, omit rates and a "motivation question" were used from a self-reported questionnaire administered before testing. Specifically, the following two data cleaning rules were applied to the data before equating: (1) exclude test takers who didn't answer at least 25% of the items on each test section and (2) exclude test takers who selected Option 3: "I gave little effort" on the self-reported survey. After data screening, the final equating sample included 4,693 test takers for the base form and 4,570 test takers for the new form.

To describe the sample used to equate Form B, Table A-6.3.1 in Appendix 6: Psychometrics includes frequency counts for the equating sample by grade level, gender, ethnicity, and first language. Also, Table 6.1 displays descriptive statistics for all twelve scores. Raw score moments for the three test scores (Math, Reading and Writing, and Language) are relatively close between the base form and the new form. However, the cross-test scores, especially Analysis of History/Social Studies, are less aligned. This is mainly due to the fact that test specifications are articulated at the test score level.

Tables A-6.5 and A-6.6 in Appendix 6: Psychometrics include raw score reliability and average CSEM values for Forms A and B. In addition, Tables A-6.7 and A-6.8 in the same appendix display the raw score reliability and average CSEM values by ethnicity, gender, and grade-level subgroups (Standard 2.11, AERA/APA/NCME, 2014). The CSEM estimates appear small and relatively constant along the score scale for all scores as expected. The reliability values are the highest, and thus SEM values are the lowest, for those scores containing the largest number of raw score points—total, followed by section, test, and finally subscores. Reliability estimates are slightly higher for males than for females and are similar across the racial/ethnic subgroups.

Table 6.1: Raw Score Distributions for Base Form (Form A) and New Form (Form B)

Score Level	Statistics N	Base Form 4,693	New Form 4,570
Math	Mean	27.9	27.5
	SD	10.1	9.9
	Skewness	0.3	0.2
	Kurtosis	−0.5	−0.5
Reading	Mean	25.9	25.6
	SD	9.1	9.9
	Skewness	0.2	0.3
	Kurtosis	−0.7	−0.8
Writing and Language	Mean	24.8	24.0
	SD	8.5	8.7
	Skewness	0.0	0.2
	Kurtosis	−0.9	−0.8
Analysis in Science	Mean	17.5	16.6
	SD	6.8	6.6
	Skewness	0.2	0.3
	Kurtosis	−0.8	−0.7
Analysis in History/ Social Studies	Mean	18.1	16.7
	SD	6.0	6.5
	Skewness	0.1	0.3
	Kurtosis	−0.7	−0.6
Command of Evidence	Mean	8.5	8.0
	SD	3.7	3.6
	Skewness	0.3	0.4
	Kurtosis	−0.6	−0.5

Table 6.1 continued on next page

Table 6.1 continued from previous page

Score Level	Statistics N	Base Form 4,693	New Form 4,570
Relevant Words in Context	Mean	10.9	10.6
	SD	3.4	3.7
	Skewness	−0.4	−0.2
	Kurtosis	−0.5	−0.7
Expression of Ideas	Mean	13.3	12.5
	SD	4.9	5.1
	Skewness	0.0	0.2
	Kurtosis	−0.9	−0.8
Standard English Conventions	Mean	11.5	11.4
	SD	4.0	4.2
	Skewness	−0.1	0.1
	Kurtosis	−0.7	−0.7
Heart of Algebra	Mean	9.5	11.2
	SD	3.7	4.4
	Skewness	0.1	−0.3
	Kurtosis	−0.5	−0.8
Problem Solving and Data Analysis	Mean	9.8	8.6
	SD	3.5	3.0
	Skewness	−0.3	0.0
	Kurtosis	−0.5	−0.3
Passport to Advanced Math	Mean	6.9	6.0
	SD	3.1	2.9
	Skewness	0.4	0.7
	Kurtosis	−0.3	0.4

Equating Methods

Several equating methods were used in the SAT equating, including linear equating and also unsmoothed and smoothed equipercentile methods. Log-linear presmoothing and cubic spline postsmoothing were considered as potential smoothing techniques for the equipercentile method. The following section provides a short introduction of the used equating and smoothing methods. Throughout the discussion, Y and X represent the base and new form, respectively.

Equation 6.2.1 displays the linear equating method (Kolen & Brennan, 2014) where the population mean and the standard deviation of the new form (X) are set to be equal to the population mean and standard deviation of the base or anchor form (Y). This method allows us to account for the differences in difficulty between two test forms along the entire score scale.

$$l_Y(x) = y = \frac{\sigma_Y}{\sigma_X}x + \left[\mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X\right] \tag{6.2.1}$$

The equipercentile equating function is another symmetric function defined by identifying scores on form X that have the same percentile ranks as the scores on form Y. Here, X and Y are assumed to be continuous random variables, uniformly distributed within each +/− 0.5 score boundary. The equipercentile equating function is described as follows

$$e_Y(x) = G^{-1}\left[F(x)\right] \tag{6.2.2}$$

where F is the cumulative distribution function of X in the population, and G^{-1} is the inverse of the cumulative distribution function *G* (Kolen & Brennan, 2014).

Cubic spline postsmoothing reduces irregularities in equipercentile equivalents by fitting a curve to the equipercentile relationship (Kolen & Brennan, 2014). This is the most widely used postsmoothing method where the cubic spline function is defined for each integer score xj as

$$\hat{d}_Y(x) = v_{0j} + v_{1j}(x - x_j)^1 + v_{2j}(x - x_j)^2 + v_{3j}(x - x_j)^3, \quad x_j \le x < x_j + 1 \tag{6.2.3}$$

The spline function is fit over a given range by minimizing the lack of smoothness subject to the following constraint (Kolen & Brennan, 2014)

$$\frac{\sum_{j=low}^{high}\left[\frac{\hat{d}_Y(x_j) - \hat{e}_Y(x_j)}{\hat{se}\left[\hat{e}_Y(x_j)\right]}\right]^2}{\left(X_{high} - X_{low} + 1\right)} \le S \tag{6.2.4}$$

where $\hat{se}[\hat{e}_Y(x_j)]$ is the analytical standard error of equipercentile equating $\hat{e}_Y(x_j)$.[2] The parameter *S* in Equation 6.2.4 controls the degree of smoothing. During the SAT field study equating a range of fixed *S* values recommended by Kolen and Brennan (2014) (*S* = 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75, and 1.00) were considered.

While cubic-spline smoothing is performed after equating, log-linear smoothing is used to reduce irregularities in the original raw score distribution before equating (von Davier, Holland, & Thayer, 2004). In this case, the moments of the resulting fitted distribution match the moments of the original unsmoothed distribution up to the number of polynomial degree specified in fitting. For the RG design, only univariate smoothing is required as no common item set (anchor set) is used. The polynomial model for the expected frequencies m_j is the following (Kolen & Brennan, 2014):

$$\log\left(m_j\right) = \omega_0 + \omega_1 x_j + \omega_2 x_j^2 + \cdots + \omega_C x_j^C \quad \left(j = 1, \ldots, J\right), \tag{6.2.5}$$

where C is the number of moments that will be matched between the raw and the fitted distributions, and the ω's are the 1, 2, . . . C parameters to be estimated by maximum likelihood estimation (ω_0 is set such that the sum of the m_j s equals the sample size). Solutions for ω = [2, 3, 4, 5, 6, 7, 8, 9] were generated for each test form separately and Akaike Information Criterion (AIC) and likelihood ratio chi-square statistic were used for selecting a final solution.

Equating Results

The characteristics of a new raw-to-scale score conversion are influenced by several factors that are predisposed before equating, such as the range and the distribution of test takers' ability, the quality of form construction, the characteristics of the base score scale, the magnitude of the difference in difficulty between the base and new forms, and the quality of form spiraling in case of the RG design. Several criteria were used to obtain the "best available" solution including:

- Prioritizing equipercentile equating over linear equating because the accuracy of the results is important along the entire score scale
- Prioritizing some degree of smoothing over a non-smoothed solution
- Avoiding more than two raw scores converting to the same scale scores, especially near the top of the score scale
- Avoiding missing scale scores (gaps), especially near the top of the score scale
- Having a close match for the scale score mean and standard deviation between the new form and the base form

In equating Form B, equipercentile solutions with cubic spline postsmoothing were chosen for all scores except for the Problem Solving and Data Analysis (PSD) subscore, in which case smoothing wasn't necessary. The S values for the final solutions ranged between 0.01 and 0.2.

[2] These standard errors reflect imprecision in the equipercentile function due to the sample of examinees. This type of error differs from the standard errors of measurement addressed in Scaling (Section 6.1) and Reliability (Section 6.4) which reflect imprecision in test scores due to the particular sample of items on the test form.

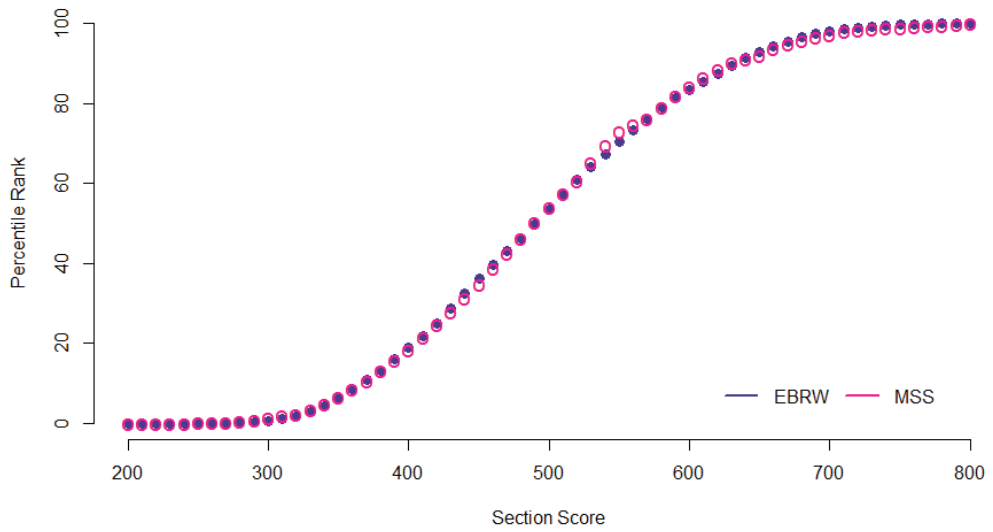
Due to the proprietary nature of the conversion tables, instead of the actual tables, a set of sample rounded conversion tables are made available for this report (see Tables A-6.3.2 through A-6.3.5 in Appendix 6: Psychometrics) for review purposes. Please note that these conversion tables don't represent any operational conversion table that was or ever will be used to score the SAT.

The purpose of equating is to make test forms interchangeable in terms of difficulty. To best recognize the effectiveness of the equating process, scale score statistics and distributions are derived and evaluated for each test form. Means and standard deviations of the base form and the new form are expected to be reasonably close. Figures A-6.8.1 through A-6.8.30 in Appendix 6: Psychometrics display test score, cross-test score, and subscore scale score statistics and distributions for both base and new forms. As shown, new form scale score statistics and distributional properties are nicely aligned with base form statistics after equating. CSEMs for scale scores are also displayed in Figures A-6.9.1 through A-6.9.12 in Appendix 6: Psychometrics, showing only a small and stable level of measurement error across the score scale.

Scale score mean and standard deviation alignment across the two section scores [Evidence-Based Reading and Writing (ERW) and the Math (MSS)] have a particular importance in SAT, as they provide succinct information for state and higher education officials. Hence, we carefully examined the behavior of SAT section scores by generating descriptive statistics and percentile rank information using the newly equated scores. Figure 6.2 displays scale score percentile ranks for the two section scores of the equated form. The scatterplot shows a very close alignment between ERW and Math section scores and a close match between means and standard deviations of the corresponding scale scores.

Data collection design and equating methods described in this section are pertinent and currently used for *operational equating* for both the SAT and the PSAT/NMSQT and PSAT 10.

Figure 6.2: Section score percentile rank for Evidence-Based Reading and Writing and Math section scores



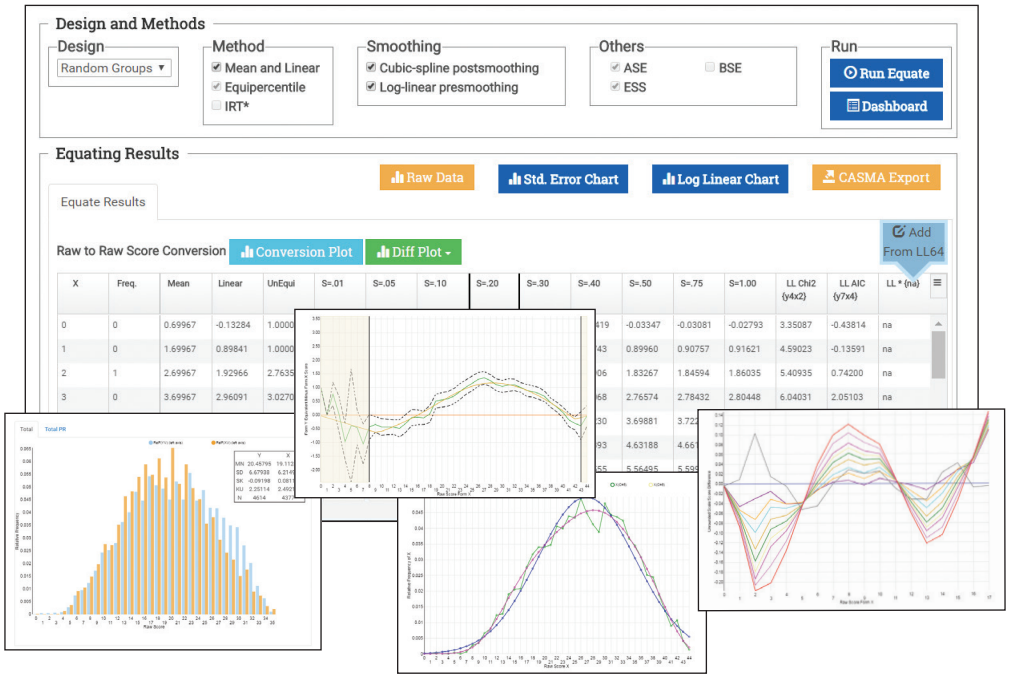
The data collection is always performed by spiraling multiple forms within each school, and linear and equipercentile equating methods are used as described previously. Log-linear and cubic spline postsmoothing are also considered as potential smoothing techniques for equipercentile equating. The sample for SAT equating is collected from designated test centers due to security and logistical reasons, while a national sample is collected to equate PSAT/NMSQT and PSAT 10 forms.

College Board's Equating System: ScorEquate

The equating of the SAT, PSAT/NMSQT, and PSAT 10 is currently supported by ScorEquate (College Board, 2015b), a new *proprietary* equating system developed by the College Board in a joint effort between its Psychometrics and Information Technology divisions. A significant component of ScorEquate is based on *Equating Recipes* (Brennan, Wang, Kim, & Seol, 2009), which is a set of open source C functions. These C routines constitute the "engine" of the equating system, which is activated by a set of customized functions via a graphical user interface.

The current version of ScorEquate is primarily tailored to the requirements of the SAT Suite of Assessments. Consequently, the system is fast and efficient, while at the same time, it is user-friendly and flexible. Several interactive graphics are built into the system to enable the user to effectively evaluate multiple competing solutions using various psychometric criteria. The system also supports multiple user roles, such as data specialists, supervisors, and primary and secondary psychometricians. The system is built to conform to the highest security requirements while users can successfully complete their tasks simultaneously and independently. The system is linked to the College Board's central databases, where the data are obtained for analysis and where approved conversion tables are saved. Figure 6.3 provides a snapshot of the current equating system.

Figure 6.3: College Board ScorEquate system



6.3 Normative Information

Normative information for the SAT Suite of Assessments was created to support appropriate interpretations of the common score scales among College Board's constituents. These statistics were constructed with the full SAT Suite of Assessments in mind, providing relevant information across the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 assessments. In line with historic practices, and best practices as outlined in the AERA/APA/NCME *Standards for Educational and Psychological Testing*, the College Board continues to provide common scale score norms associated with all reported scores for all three assessments (AERA/APA/NCME, 2014). In this section, the norming methods used for the SAT Suite are defined in terms of their statistics, sample data, target populations, accuracy evaluations, and plans for periodic updating.

Norms

According to AERA/APA/NCME Standard 5.10, the statistics used to characterize examinee groups should be clearly defined and should support the intended use or interpretation of the reported common scale scores. To this end, definitions are incorporated into the normative information. In addition to percentile rank norms (for each common scale score point on each of the 15 new SAT, PSAT/NMSQT and PSAT 10, and 14 new PSAT 8/9 reported scores), reference group averages have been provided to allow schools, districts, and states to compare themselves to the target populations in terms of aggregate statistics. Moreover, measures of accuracy of the reported normative information have been provided for each common scale score in each reference group.

The percentile rank statistic is the basis of score interpretations with respect to reference groups of interest. Percentile ranks, used as the reported common score scale normative information, are defined in a manner that is consistent with the scaling and equating definitions, in that they indicate the percentage of students at or below a given scale score point:

$$PR_{SC_j} = \frac{\left(\sum_{q=1}^{j-1} f_q\right) + \frac{f_j}{2}}{N} \times 100\%$$

where the percentile rank (*PR*) for a scale score (sc_j) is given by the quotient of the sum of the cumulative frequencies below the considered score (f_q, $q = 1, \ldots, j-1$), plus half the frequency at the score point (f_j), and the total sample size (N). The percentile rank can also be expressed as the mean of test taker contributions, $PR_{sc_j} = \frac{\sum_t PR'_{sc_{jt}}}{N}$ where $PR'_{sc_{jt}} = 1 \times 100$ if $sc_t < sc_j$, $PR'_{sc_{jt}} = \frac{1}{2} \times 100$ if $sc_t = sc_j$, and $PR'_{sc_{jt}} = 0 \times 100$ if $sc_t > sc_j$.

Similar to the normative information provided for the old SAT, score norms for the SAT Suite are reported in rounded integer form. Percentile ranks that round to 0 (unrounded percentile rank values < 0.50) are denoted as "1−", while percentile ranks that round to 100 (unrounded percentile rank values >= 99.50) are denoted as "99+." In addition to scale score specific norms, scale score means (MEAN) and standard deviations (SD) are produced (also rounded to the nearest integer) for the target population groups.

Reference Groups

In keeping with AERA/APA/NCME Standards 5.8 and 5.9, reported score norms should reflect the examinee populations against which the test takers would seek to compare themselves. The College Board provides two target populations, which allow test takers to compare their performance to that of similar cohort students and also against the performance of a more diverse national group. Test takers aren't the only users of the reported common scale scores. Additional constituents, such as college admission offices, high school counselors, state school districts, etc., make use of the reported common scale scores for a variety of purposes. Thus, the College Board made efforts to identify and provide the aforementioned two reference groups that would be generalizable and of interest to all consumers of the common scale scores.

Standards 5.8 and 5.9 also state that definitions and details regarding the target populations should be provided for a clear and transparent understanding of which students comprise the reference groups. Specific information regarding the sampling/construction of the relevant target population samples, participation rates, descriptive statistics, etc., has been provided throughout this technical documentation. Specific reference group common scale score performance moments (means and standard deviations) are listed in the following text.

SAT

The two SAT target populations for which normative information is provided are a nationally representative combined 11th- and 12th-grade group, and a "user" group of representative College Board graduating high school seniors who have taken their last SAT assessment in the 11th or 12th grade (referred to as the College Board cohort).

The estimated nationally representative normative group for the SAT follows a demographic composition consistent with the national demographics that are captured by NCES's Common Core of Data and the Private School Universe Survey. New SAT common scale score norms were established based on the test takers from the scaling study who took the two forms. The same motivation screenings and weighting methodology used in scaling were used to aggregate test taker performances to the nationally representative group levels (Haberman, 1984, 2015). See Section 6.1: Scaling for sampling, examinee, screening, and weighting technical details. Equating and scaling procedures were used to obtain scale scores for the screened students who took the two scaling study forms. The scaling study examinees were combined with screened examinees from the pre-2016 SAT and the redesigned SAT concordance study (see Section 7.3: The Relationship Between the SAT and Other Similar Assessments), which collectively provided the data for the norms estimates of the SAT nationally representative group.

The first set of SAT User group normative information was estimated based on a sample weighted to follow a demographic composition equivalent to the average proportions of three domestic graduating cohorts (2012–2014; retaining students who have last taken the SAT as an 11th or 12th grader) based on concordance study examinee performances (see Section 7.3: The Relationship Between the SAT and Other Similar Assessments). Future SAT College Board user group norms won't be reestimated on field trial data but rather on the operational performances of the College Board graduating cohorts (see Updating of Norms in the following text).

Table 6.2 summarizes the means and standard deviations for the nationally representative and College Board user reference groups for each of the 15 reported SAT common scale

scores (total score, 2 sections scores, 5 test scores, and 7 subscores). Normative information was based on N = 8,677 examinees in the user group and N = 19,048 in the nationally representative sample (note that weighting preserved these sample sizes and thus the effective postweighting sample size reflects the number of collected examinee responses). The means and standard deviations in Table 6.2, as well as the nationally representative and College Board user norms for the SAT, are provided for the combined 11th- and 12th-grade group, and not separately by grade. Grade-specific norms won't be provided for either the nationally representative or the College Board user SAT normative samples.

Table 6.2: SAT Target Population Common Scale Score Means (MEAN) and Standard Deviations (SD)

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Total	Total	1020	194	1083	193
Section	ERW	510	102	543	102
	MSS	510	103	541	103
Test	R	25	5	27	5
	WL	26	5	27	5
	MTS	25	5	27	5
	HSS	26	5	27	5
	SCI	25	5	27	5
Subscore	COE	8	3	9	3
	WIC	8	3	9	3
	SEC	8	3	9	3
	EOI	8	3	9	3
	HOA	8	3	9	3
	PSD	8	3	9	3
	PAM	8	3	9	3

Note: For SAT score details, see Section 5.1 (Scoring Procedures).

The nationally representative and College Board user scale score norms (percentile ranks) are provided in Tables A-6.4.1 through A-6.4.4 in Appendix 6: Psychometrics. Both the nationally representative normative information and the College Board user group norms are provided to test takers on their printed score reports and are made available through the College Board's online score reporting portal.

PSAT/NMSQT and PSAT 10

Unlike the SAT normative information, the PSAT/NMSQT and PSAT 10 norms are provided conditionally by grade. Both an 11th-and a 10th-grade PSAT/NMSQT and PSAT 10 set of norms were created and will be maintained. Furthermore, like the SAT norms, the PSAT/NMSQT and PSAT 10 grade-specific normative information are also provided for the two previously discussed reference groups.

The grade-specific estimated nationally representative normative group for the PSAT/NMSQT and PSAT 10 has a demographic composition consistent with the national demographics that are captured by NCES's Common Core of Data and the Private School Universe Survey, as was the case with the SAT. The same data preparation methodology used in scaling was used to aggregate test taker performances to the nationally representative group levels (see Section 7.3: How Does the SAT Relate to PSAT-Related Assessments?).

The PSAT/NMSQT and PSAT 10 user group normative information was based on a sample of 10th and 11th graders who took either or both the PSAT/NMSQT in fall 2015 and the PSAT 10 in the spring administration in 2016. Future PSAT/NMSQT and PSAT 10 user group norms will continue to be established in this manner with a variation on examinee selection (see Updating of Norms in the following text).

Tables 6.3 and 6.4 summarize the means and standard deviations for the nationally representative and College Board user reference groups for each of the 15 reported PSAT/NMSQT and PSAT 10 common scale scores by the 11th and 10th grade levels, respectively. Normative information was based on N = 2,209 11th and N = 8,984 10th graders in the nationally representative sample (weighting to preserve the effective sample size), and N = 1,781,802 11th and N = 2,151,254 10th grade examinees in the user group.

Table 6.3: PSAT/NMSQT and PSAT 10 11th-Grade Target Population Common Scale Score Means (MEAN) and Standard Deviations (SD)

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Total	Total	969	168	1009	193
Section	ERW	480	92	507	104
	MSS	489	88	502	103
Test	R	24	5	25	5
	WL	24	5	25	6
	MTS	24	4	25	5
	HSS	24	5	25	5
	SCI	24	5	25	5

Table 6.3 continued on next page

Table 6.3 continued from previous page

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Subscore	COE	8	2	9	3
	WIC	8	3	9	3
	SEC	8	2	9	3
	EOI	8	2	9	3
	HOA	8	3	9	3
	PSD	8	2	9	3
	PAM	8	3	9	3

Note: For PSAT/NMSQT and PSAT 10 score details, see Section 5.1 (Scoring Procedures).

Table 6.4: PSAT/NMSQT and PSAT 10 10th-Grade Target Population Common Scale Score Means (MEAN) and Standard Deviations (SD)

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Total	Total	939	170	933	176
Section	ERW	468	94	468	98
	MSS	470	88	464	92
Test	R	24	5	24	5
	WL	23	5	23	5
	MTS	24	4	23	5
	HSS	23	5	23	5
	SCI	24	5	24	5
Subscore	COE	8	2	8	3
	WIC	8	3	8	3
	SEC	8	3	8	3
	EOI	8	2	8	2
	HOA	8	3	8	3
	PSD	8	2	8	2
	PAM	8	3	8	3

Note: For PSAT/NMSQT and PSAT 10 score details, see Section 5.1 (Scoring Procedures).

The nationally representative and College Board user scale score norms (percentile ranks) for the PSAT/NMSQT and PSAT 10 11th and 10th grades, respectively, are provided in Tables A-6.4.5 through A-6.4.8 and Tables A-6.4.9 through A-6.4.12 in Appendix 6: Psychometrics.

The PSAT/NMSQT (administered in the fall) score reports will make use of the same normative information as the PSAT 10 (administered in the spring).

PSAT 8/9

PSAT 8/9 normative information is established using the same methodology as for PSAT/NMSQT and PSAT 10. Again, PSAT 8/9 norms are provided conditionally by 9th- and 8th-grade examinees for the nationally representative and College Board user reference groups.

The grade-specific estimated nationally representative normative group for the PSAT 8/9 has a demographic composition consistent with the national demographics that are captured by NCES's *Common Core of Data and the Private School Universe Survey*. Section 7.3 (The Relationship Between the SAT and Other Similar Assessments) describes the sampling, examinees, screening, and weighting procedures.

PSAT 8/9 user group normative information was based on a sample of 9th and 8th graders who took the PSAT 8/9 in fall 2015 and/or spring 2016.

Tables 6.5 and 6.6 summarize the means and standard deviations for the nationally representative and College Board user reference groups for each of the 14 reported PSAT 8/9 common scale scores by the two grade levels, respectively (the PSAT 8/9 doesn't have a Passport to Advanced Mathematics, or PAM, subscore as compared to the PSAT 10 and the SAT). Normative information was based on N = 17,091 9th graders and N = 7,263 8th graders in the nationally representative sample (weighting to preserve the effective sample size), and N = 509,680 9th graders and N = 360,582 8th graders in the user group.

Table 6.5: PSAT 8/9 9th-Grade Target Population Common Scale Score Means (MEAN) and Standard Deviations (SD)

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Total	Total	889	154	868	165
Section	ERW	446	85	438	90
	MSS	443	81	430	87
Test	R	23	4	22	5
	WL	22	5	21	5
	MTS	22	4	21	4
	HSS	22	5	22	5
	SCI	23	4	22	5

Table 6.5 continued on next page

Table 6.5 continued from previous page

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Subscore	COE	8	3	8	3
	WIC	8	3	8	3
	SEC	8	3	8	3
	EOI	8	3	8	3
	HOA	8	3	8	3
	PSD	8	3	8	3

Note: For PSAT 8/9 score details, see Section 5.1 (Scoring Procedures).

Table 6.6: PSAT 8/9 8th-Grade Target Population Common Scale Score Means (MEAN) and Standard Deviations (SD)

		National		User	
Level	**Score**	**MEAN**	**SD**	**MEAN**	**SD**
Total	Total	835	137	808	150
Section	ERW	422	76	406	84
	MSS	414	75	401	79
Test	R	21	4	21	4
	WL	21	4	20	5
	MTS	21	4	20	4
	HSS	21	4	20	5
	SCI	21	4	20	5
Subscore	COE	7	2	7	3
	WIC	7	3	7	3
	SEC	7	3	7	3
	EOI	7	2	7	3
	HOA	7	2	7	2
	PSD	7	2	7	3

Note: For PSAT 8/9 score details, see Section 5.1 (Scoring Procedures).

The nationally representative and College Board user scale score norms (percentile ranks) for the PSAT 8/9 9th and 8th grades are provided in Tables A-6.4.13 to A-6.4.16 and Tables A-6.4.17 to A-6.4.20, respectively, in Appendix 6: Psychometrics.

Special Groups

Currently, the College Board doesn't produce special subgroup normative information, such as for gender, race/ethnicity, or language subgroups. During the scaling study, the College Board included test takers with disabilities who had requested testing accommodations. The Services for Students with Disabilities (SSD) office matched those normally provided accommodations in regular SAT administrations. Although these data are available for future study, examinees who tested under accommodations were excluded from the computation of normative information due to the differential administration mode associated with their requested testing accommodations.

Precision

Normative information, both at the cohort and the nationally representative level, is monitored on an ongoing basis through the College Board's standards of test evaluation and maintenance. For example, normative information of the SAT user group is updated periodically to assure the relevance and accuracy of constituents' comparisons with the reference groups of interest.

Multiple analytical and resampling statistics for computing standard errors of the percentile rank estimates were created to summarize the sampling stability (i.e., precision) of the reported norms.[3] Standard errors are derived via several approaches, including analytical and bootstrap resampling. For use in technical documentation, the analytical standard error of the scale score associated percentile rank is reported as the preferred measure of accuracy for the normative information.

For operational data, as is the case in the estimation of the College Board PSAT-related user group normative information, conditional analytical standard errors can be computed directly from the data by:

$$SE_{sc_j} = \sqrt{\frac{SD^2_{PR'_{sc_j}}}{N}},$$

where for any scale score (sc_j) with a specific percentile rank (PR_{sc_j}), the conditional standard error of that percentile rank is estimated as the square root of the estimated variance of test takers' contributions to that percentile rank ($PR'_{sc_{jt}}$) conditional on the scale score point divided by the total sample size.

Because the SAT user norms and all of the nationally representative norms are based on weighted data, conditional percentile rank standard errors, based on field trial data that is weighted, need to take the weighting model under consideration. The conditional standard errors of percentile ranks computed via weighted data are estimated based on the weighted

[3] The standard errors for norms reflect errors due to samples of examinees, which are different from the standard errors of measurement discussed in the scaling and reliability sections that reflect errors due to samples of test items.

average of the squared residuals of the differences of the t^{th} test taker's contributions to a percentile rank ($PR'_{sc_{jt}}$) from a regression-based prediction of the weighting model (Haberman, 1984).

The bootstrap conditional standard error is also computed for comparative purposes. This standard error can be applied with either field trial and weighted or operational data and is defined as the standard deviation of the percentile ranks within a scale score (sc_j) across bootstrap iterations (B).

Standard errors for percentile ranks tend to vary across score scales, being at their largest values around the middle of the scale score range where most data are, and nearly zero at the lowest and highest scores.

Updating of Norms

In order to assure continued accuracy of norms and score interpretability, and in keeping with AERA/APA/NCME Standard 5.11, common scale score norms need to be reestablished with appropriate frequency. During the initial launch year of the SAT Suite of Assessments, great efforts have been made to establish normative information that meet all industry and College Board standards. Within the delineated multiyear plan of implementing the SAT Suite of Assessments, the College Board has also provided a four-year trajectory for building steady-state normative information and a yearly updating schedule to assure the relevance of the normative information.

Nationally representative normative information for all three exams, SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9, will be monitored and updated as needed. Using a special field study design, or by appropriately weighting nationally collected administration data, the nationally representative percentile rank information will reflect the current state of the nation's makeup of grade appropriate examinees (as this isn't expected to change rapidly, it will be updated with several year intervals).

The new SAT user group scale score norms will be updated based on averaged information. The first set of normative information was established based on concordance study data. These norms will be replaced with the first new SAT cohort norms computed in July 2017. Subsequent updating of the norms based on the SAT College Board user group cohort will begin by compounding cohorts (2017, 2017–2018, and 2017–2019) until a three-year total is reached. In the steady state administrations of the SAT, user group norms will be updated on a yearly basis using a three-year rolling average, dropping the first year when adding new cohort information (i.e., 2020 norms would be based on 2018–2020, 2021 norms on 2019–2021 cohorts, and so forth).

Because the PSAT 10 and PSAT 8/9 launched in 2015, prior to the SAT, the PSAT/NMSQT and PSAT 10, and PSAT 8/9 user group normative information has already been established based on fall 2015 and spring 2016 examinees. Both the PSAT 10 (which includes the PSAT/NMSQT) and the PSAT 8/9 user group norms will be updated yearly following the same SAT User group updating methodology discussed earlier in this chapter, but will use all examinees with reported scores, rather than cohort type data. This means that the three-year rolling average norms for the PSAT 10 and PSAT 8/9 are equivalent to norms based on reportable scores from six administration windows, the fall and spring of three years.

6.4 Reliability

Reliability is a measure of consistency in test takers' observed scores. As discussed in Section 1.1 of this manual, consistency in scores across instances of the test procedure is one component to ensure that scores are valid for their intended uses. Test takers' observed scores may vary for many reasons. This variance can occur, for example, if the test is administered at two different points in time, across different forms of a test, or due to changes in test administration or scoring conditions. There are many different methods to estimate reliability coefficients, including those based on Generalizability Theory, Classical Test Theory, and Structural Equation Modeling. For the SAT Suite, Kuder-Richardson 20 (KR20) and stratified coefficient alpha are calculated for raw scores and the compound binomial model is used for scale scores. Reliability estimates range from 0 to 1, with values near 1 indicating more consistency and values near 0 indicating little to no consistency. In this section, reliability and standard error of measurement (*SEM*) estimates are described that are most appropriate for the scores of the new tests in the SAT Suite of Assessments (AERA/APA/NCME, 2014).

Reliability Indices

KR20 is an estimate of internal consistency, specifically useful when all items are dichotomously scored and the parts are essentially tau equivalent (Haertel, 2006). Stratified alpha is the most appropriate reliability estimate of a composite score or when the assumption of congeneric parts is not met (Haertel, 2006). The number of strata for each score tier may be determined by the number of passages contributing to the score tier (e.g., test scores and subscores related to the Reading Test and the Writing and Language Test). It can also be determined by the number of timed sections contributing to the same score tier (e.g., Math test score, subscores involving items from different timed sections, and cross-test scores). The compound binomial model is used to calculate scale score conditional standard error estimates. See Section 6.1: Scaling for more information on how this model is applied.

Kuder-Richardson 20

Kuder-Richardson 20 (KR20) reliability estimate targets are used in the test specifications for the SAT Suite of Assessments. Thus, the estimates observed from each administration can be compared to the target KR20 values to evaluate the reliability aspect of this test administration.

$$KR20_X = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \varphi_i (1-\varphi_i)}{\sigma_X^2}\right)$$

where

n represents the number of parts or items within the test,

φ_i is the item difficulty (p-value, see Exhibit A-5.1 in Appendix 5) for item i, and

σ_x^2 is the population variance of the total test score that is estimated as $\sqrt{\frac{\sum_t \left(X_t - \overline{X}\right)^2}{N}}$

[N is the number of test takers and differs from n, which is the number of parts or items (Haertel,2006)].

The KR20 reliability estimate is equivalent to coefficient alpha when the parts within the test are considered items and are scored dichotomously (0 or 1), as in the case for the assessments considered here.

Stratified Alpha

Stratified alpha is the most appropriate reliability estimate of a composite score or when the assumption of congeneric parts, or strata, is not met (Haertel, 2006).

$$SA_X = 1 - \frac{\sum \sigma^2_{X_i}\left(1 - \rho_{XX'_i}\right)}{\sigma^2_X}$$

where *i* represents the number of strata within the test, $\rho_{xx'_i}$ is the reliability coefficient for strata *i*, $\sigma^2_{x_i}$ is the population variance of the strata *i* score that is estimated as $\sqrt{\frac{\sum_t \left(X_{it} - \overline{X_i}\right)^2}{N}}$ where *t* represents test taker, and σ^2_x is the population variance of the total test score that was defined previously for $KR20_x$ (Haertel, 2006). For the SAT Suite of Assessments, the $\rho_{xx'_i}$ is replaced with $KR20_i$, the KR20 reliability estimate for part i.

The number of strata for each score tier was determined by the number of passages contributing to the score tier [e.g., test scores and subscores related to Reading (R) and Writing and Language (WL)] or the number of timed sections contributing to the same score tier [e.g., Math test score, subscores involving items from different timed sections (i.e., Math No Calculator (MNC) and Math with Calculator (MWC), and cross-test scores)]. See Table 6.7 for descriptions of the number and type of strata for the various raw and scale scores.

Standard Errors of Measurement (SEM) and Conditional Standard Errors of Measurement (CSEM)

The standard error of measurement (SEM) provides an estimate of the amount of error or inconsistency in observed test scores.[4] SEM is calculated as

$$SEM = SD_X\sqrt{1 - \rho_{XX'}}$$

where SD_X is the estimated standard deviation of score X and $\rho_{XX'}$ is the reliability of score X (Crocker & Algina, 1986).

SEM is calculated for each of the raw score reliability estimates, KR20 and stratified alpha, by substituting $\rho_{xx'}$ with the appropriate estimate of reliability for score X.

Conditional standard errors of measurement (*CSEM*) estimate the amount of error or inconsistency in observed test scores for a particular true score. Section 6.1 describes how the scales were developed to have approximately equal standard errors of measurement for all true scale scores, that is, constant conditional standard errors of measurement.

[4] This differs from the standard errors discussed in the previous chapter in that those were standard errors due to sampling that reflect samples of examinees rather than standard errors of measurement that reflect samples of test items.

Table 6.7: Description of the Number and Type of Strata Used in Stratified Reliability

	Scores	Source of Items	Type of Strata	Number of Strata	Number of Raw Score Points
Test Scores	**Evidence-Based Reading (R)**	All from R	Passage	5	52
	Writing and Language (WL)	All from WL	Passage	4	44
	Math	All from MWC and MNC	Calculator usage	2	58
Math Timed Sections	**Math No Calculator (MNC)**	All from MNC		1	38
	Math with Calculator (MWC)	All from MWC		1	20
Cross-Test Scores	**Analysis in Science**	6 from WL, 21 from R, 9 from Math	Passage	3	35
	Analysis in History/ Social Studies	6 from WL, 21 from R, 9 from Math	Passage	3	35
Subscores	**Heart of Algebra**	13 from MWC, 8 from MNC	Calculator usage	2	19
	Passport to Advanced Math	7 from MWC, 9 from MNC	Calculator usage	2	16
	Problem Solving and Data Analysis	MWC		1	17
	Expression of Ideas	WL	Passage	4	24
	Standard English Conventions	WL	Passage	4	20
	Relevant Words In Context	8 from WL, 10 from R	R/WL	2	18
	Command of Evidence	8 from WL, 10 from R	R/WL	2	18
Section Scores	**Evidence-Based Reading and Writing**	All from WL and R			96
	Math	All from MWC and MNC			58
Composite		**Sum of Section Scores**			**154**

For scale scores, reliability estimates were obtained from the average squared CSEM across all possible scale scores using the following equation:

$$\text{Reliability}_{sc} = \rho_{sc,sc'} = 1 - \frac{MS(CSEM)_{sc}}{SD_{sc}^2},$$

where SD_{sc}^2 is the estimated variance of the scale scores. $MS(CSEM)_{sc}$ is the mean squared CSEM and was obtained by calculating a weighted average of the squared CSEMs for the scales directly established.

For the scores that were mathematically derived, including Math Test (MTS), Evidence-Based Reading and Writing (ERW), and total scores, the following equations were used to compute the root mean square CSEMs, RMS(CSEM), as previously described in Section 6.1:

$$RMS(CSEM)_{MTS} = \sqrt{\frac{MS(CSEM)_{MSS}}{20^2}}$$

$$RMS(CSEM)_{ERW} = \sqrt{MS\left(CSEM\right)_{R} \cdot 10^2 + MS\left(CSEM\right)_{WL} \cdot 10^2}$$

$$RMS(CSEM)_{Total} = \sqrt{MS\left(CSEM\right)_{ERW} + MS\left(CSEM\right)_{MSS}}$$

where MSS is the Math section score. These equations are based on the assumption that measurement errors from the contributing scores (e.g., Reading and Writing and Language) are independent.

Reliability, *SEM*, and *RMS*(*CSEM*) values for raw scores and/or scale scores for an SAT form are in Appendix 6: Psychometrics. The results are based on unweighted data, and response records were excluded if the test taker didn't respond to at least one item in each timed section or if the test taker's score was on a security hold.

Raw and scale score reliability and average *CSEM* values can be found in Table A-6.5 and Table A-6.6, respectively, in Appendix 6: Psychometrics. See Figures A-6.9 to A-6.15 for graphical representation of the conditional standard error of measurement for each scale score for test, cross-test, and subscores. See Tables A-6.7 and A-6.8 in the same appendix for the raw and scale score reliability and average *CSEM* values by ethnicity, gender, and grade-level subgroups (AERA/APA/NCME, 2014).

The results show high consistency in the reliability and *SEM* values between the KR20 and stratified coefficient results, including the results for the subgroups. As would be expected, the reliability values are the highest, and thus *SEM* values the lowest, for those scores containing the largest number of raw score points—total, followed by section scores, test scores, and finally subscores. Reliability estimates are slightly higher for males than for females and are similar across the racial/ethnic subgroups.

6.5 Psychometric Applications

In addition to the scaling, equating, norming, and reliability analyses described in the previous psychometric sections, there are psychometric analyses that have particular applications to the overall processes of item development, test form assembly, and test score evaluation. These additional psychometric analyses are briefly summarized here, pointing out how item analyses results are used to evaluate test forms with respect to statistical targets; test score distributions and moments are used to evaluate the overall functioning of the SAT Suite

scores for populations and subgroups; completion percentages are used to evaluate the speededness of test forms; and correlations are used to evaluate test score relationships.

Statistical Targets

The statistical targets include specifications of the average item difficulty estimates, the lower and upper bounds for item discrimination, and the lower and upper bounds for reliability. The consistency of item difficulty across forms is important to ensure the assessment contains items that are appropriate (not too difficult and not too easy) for the intended test takers. The item discrimination values are important for monitoring the ability of the items to accurately distinguish higher performers from lower performers. The reliability estimates of subsequent test forms are monitored to ensure consistency in scores across forms (see Section 6.4). Checking the operational data to ensure these targets are met consistently across forms is important for quality control purposes.

The overall functioning of the test form and items are considered in several ways. For the SAT Suite, the first four moments of the raw and scale score distributions are examined, the correlations among various reported scores are calculated, and the item completion rates and form speededness statistics are considered, along with the reliability estimates for the raw and reported scale scores. Several of these statistics are also examined for gender, a number of ethnic subgroups, and grade levels in order to assess the equivalent functioning of the form for various subgroups. These include the moments and standardized mean differences between the specified groups, correlations, reliability estimates (see Section 6.4), and speededness statistics.

Results of these analyses for several SAT, PSAT/NMSQT and PSAT10, and PSAT 8/9 forms are included in Appendix 6: Psychometrics. The results are based on unweighted data, and response records were excluded if the test taker didn't respond to at least one item in each timed section or if the test taker's score was under a security hold.

Intercorrelations

Relationships among test scores are useful for understanding psychometric qualities such as what the scores measure and also for the implications of using and applying scores for different purposes (e.g., predictive relationships, uses of subscores vs. total scores). The Pearson product moment correlation coefficient provides an evaluation of the pairwise linear relationship between the total score, section scores, test scores, cross-test scores, and subscores. The disattenuated, or true score, correlations are the correlations after correcting for attenuation between the two scores, specifically after removing the weakening effects of measurement error to the correlations. The formulas for calculating the Pearson correlations and disattenuated, or true score, correlations are provided in the following text.

Pearson Product Moment Correlation Coefficient

The Pearson product moment correlation coefficient is calculated as

$$\rho_{XY} = \frac{\sum Z_X Z_Y}{N}$$

where Z_X and Z_Y represent Z-scores (e.g., Section 5.2) of observed scores X and Y, respectively, and N represents the number of test takers (Crocker & Algina, 1986).

Disattenuated Correlations/True Score Correlations

The disattenuated correlations are calculated as

$$\rho_T = \frac{\rho_{XY}}{\sqrt{SA_X SA_Y}}$$

where ρ_{XY} is the correlation between observed scores *X* and *Y*, while SA_X and SA_Y represent the stratified alpha reliability of score *X* and *Y*, respectively (Schumacker & Muchinsky, 1996). See Section 6.4 for more information about stratified alpha.

Table A-6.9 in Appendix 6: Psychometrics provides observed and true score correlations among the raw scores, and Table A-6.12 in the same appendix provides scale score correlations for the total group. The correlations above the diagonal represent the true score correlations, while the correlations below the diagonal represent the observed score correlations.

The results show relatively high correlations among similar content scores [e.g., Reading with the Reading-related subscores (COE and WIC)] and lower correlations among dissimilar content scores [e.g., Reading with Math-related subscores (HOA, PSD, PAM)]. The largest observed score correlations were those between the Writing and Language Test score and the Writing-related subscores (EOI and SEC), as well as between the Evidence-Based Reading and Writing section score and the Reading Test and Writing and Language Test scores.

Moments and Score Distributions

Test taker performance is described using the first four moments for all score tiers (raw and scaled). The mean, standard deviation, skewness, and kurtosis provide a description of the distribution of scores, whereby skewness indicates the extent to which the distribution is symmetrical, and kurtosis indicates whether the distribution is peaked or flat relative to a normal distribution. The test taker performance for each subgroup is described using the first four moments of the subgroup-specific data for all score tiers. Standardized mean differences are also included to provide a value of the magnitude of difference in performance between the groups, an important aspect of fairness evaluations. The formula for computing standardized mean difference is

$$Std.\ Diff. = \frac{\overline{X}_f - \overline{X}_r}{SD_{X,P}}$$

where $\overline{X}_f$ and $\overline{X}_r$ represent estimated mean raw scores for the focal group and reference group (white or male), respectively. $SD_{X,P}$ represents the pooled standard deviation (Cohen, 1988):

$$SD_{X,P} = \sqrt{\frac{\left(n_f - 1\right) SD^2_{X_f} + \left(n_r - 1\right) SD^2_{X_r}}{n_f + n_r - 2}}$$

where the *n*s and SD^2_Xs are the sample sizes and estimated variances of *X* for the focal and reference groups. The standardized mean differences for scale scores are computed by replacing the raw scores denoted by *X* with scale scores.

Test taker performance is also represented by raw and scale score frequency distributions.

See Tables A-6.10 to A-6.11 and Tables A-6.13 to A-6.14 in Appendix 6: Psychometrics for raw and scale score moments and standardized mean differences between subgroups for several SAT, PSAT/NMSQT and PSAT10, and PSAT 8/9 forms. Figures A-6.16 through A-6.29 in the same appendix show raw and scale score distributions for the same set of forms.

The results show that within each subgroup, the average scale scores were similar within each score tier. That is, within each subgroup, test takers performed similarly on all of the test scores, performed similarly on all of the subscores, and performed similarly in terms of the section scores. The largest standardized mean differences are between the white and black test takers.

Item Completion Rates and Form Speededness

Item completion rates reflect the percentage of test takers reaching an item within each timed section, which discusses how realistic it is for test takers to complete items in various sections within the allocated time limits. A reached item is one that has at least one subsequent item within the same timed section with a response. Conversely, a nonreached item is one that has no subsequent item within the same timed section with a response. Test form speededness is evaluated by examining:

- The number of items reached by at least 80% of the test takers.
- The percentage of test takers completing at least 75% of each timed section.
- The mean and standard deviation of the number of items not reached.

Seventy-five percent of a timed section is determined by the ceiling of 75% of the section length. For example, if a section has 47 items, the statistic is calculated as the percentage of test takers completing 36 or more items in the section. The degree of speededness of a test is considered negligible when 80% of the students reach the last item and when all students reach at least 75% of the questions (van der Linden, 2011). However, judgments of appropriateness of timing should be made using all relevant data.

Tables A-6.15 and A-6.16 in Appendix 6: Psychometrics provide item completion rates and section-level speededness statistics for the total group, while Table A-6.17 provides section-level statistics for selected subgroups.

The results show that in general for SAT, somewhere between 95% and 100% of subgroup members completed at least 75% of a given test section. For PSAT/NMSQT and PSAT 10, between 89% and 99% of test takers in each subgroup completed at least 75% of each section. For PSAT 8/9, Form 1 was an atypical and more difficult form and doesn't represent PSAT 8/9 forms in terms of completion rates. Form 2 is a typical form representing the completion rates for PSAT 8/9, and it had between 91% and 100% of subgroup members completing at least 75% of each test section. Additionally, generally for all assessments, 80% of the test takers, including all subgroups, completed the last item in the Reading and the Writing and Language timed sections. However, the results for the total group and the subgroups show that not all of these test takers are able to complete the last 1–4 items in the two timed Math sections. It is important to note that these last items in the Math sections are student-produced response (SPR) items and may be distorting the speededness data because of an interaction with response format and item location. For example, if the SPR items were placed at an earlier point in the test, and not at the end, it's likely that completion rates would be higher for the Math sections, and omit rates for the SPR items would be higher than for multiple-choice items.

CHAPTER 7

Validity

This final section covers validity. As we have hopefully emphasized throughout this manual, ensuring that test score interpretations are supported by strong validity evidence for their intended uses is a key component of quality assessments. As such, a commitment to matters of validity is of paramount importance to us as we develop, administer, and score the SAT Suite of Assessments, as well as any other College Board assessment. In many ways, every chapter of this manual covers validity, as all of the procedures described in the previous sections attempt to ensure that the SAT Suite produces scores that are valid for their intended uses. We have placed a deeper examination of issues of validity here at the end of this manual, as we believe that by first acquiring an understanding of the many processes involved in creating, administering, and scoring the SAT Suite, and understanding how and why we interpret those scores for their intended uses, one can more completely comprehend the steps taken toward establishing sound validity evidence.

Our examination of validity as it relates to the SAT Suite 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. We then shift focus to the test itself. The SAT Suite was redesigned with an eye toward validity, and Section 7.2 presents the evidentiary foundations behind the test content in the SAT Suite. After establishing the evidentiary foundations of the SAT Suite, Section 7.3 summarizes the concordance between the scores on the new and the old SAT, as well as the new PSAT/NMSQT to the old PSAT/NMSQT, while the second part of the section describes the vertical scaling methods used to link the new SAT to the new PSAT/NMSQT. We then turn our attention outward and look at how our scores support their intended uses and interpretations. In particular, we look at the test scores as predictors of postsecondary success and as indicators of college readiness. Section 7.4 examines the relationship SAT scores have with first-year grade point average (FYGPA) and course grades in specific subjects in college. Lastly, Section 7.5 describes the process for creating and validating the SAT College and Career Readiness Benchmarks, which communicate a student's likelihood of postsecondary success, while Section 7.6 describes how the benchmarks for the earlier assessments in the SAT Suite were derived from the SAT benchmarks.

7.1 Introduction to Validity as a Concept

Gathering validity evidence is one of the most important steps in creating and understanding test scores and their uses. As per the AERA/APA/NCME Standards, validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA/APA/NCME, 2014, p. 11). This means that if a test score has multiple uses and interpretations (e.g., admission to an institution *and* placement into coursework), each distinct use and interpretation must be validated. Note that validity is not a property of the assessment itself but rather refers to the interpretation of test scores for a specific use.

The goal in test score validation is to develop a logical rationale for the proposed uses of test scores and then to find evidence to support (or refute) those uses (Kane, 1992, 2006,

2013; Sireci, 2013). For the SAT Suite, this validity argument would include the evidence from new studies and some previously reported research on the utility of scores for the purposes of evaluating and monitoring students' college and career readiness and making college admission decisions.

The validity evidence gathering process can be thought of as an iterative, never-ending process to help improve the assessment for each use (Kane, 2013). The more evidence collected supporting a score's particular use, the stronger the argument for that particular interpretation of the test score. Validity evidence for the SAT Suite is gathered in multiple ways. Evidence includes (but isn't limited to) consideration of the content of the assessment (e.g., subject areas covered and the format of the items), the internal structure of the assessment (e.g., psychometric properties), and how test scores relate to other variables (e.g., predictive relationships). This chapter outlines and discusses the validity evidence gathered to support the proposed uses and interpretations of the SAT Suite.

7.2 Content-Oriented Validity Evidence

What Does the SAT Suite Measure?

The new SAT Suite is intended to assess the skills, knowledge, and understandings that matter most for college and career readiness and, in turn, the resulting scores from the assessments are intended to be interpreted in regard to a student's readiness for college and career training programs. According to 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). This section describes the evidentiary foundations for the decisions made about the content included on the SAT Suite's Evidence-Based Reading and Writing Tests, the Essay, and the Math Tests.

Evidentiary Foundation for the SAT Suite's Evidence-Based Reading and Writing Tests and the SAT Essay

Two tests comprise the SAT Suite's Evidence-Based Reading and Writing (ERW) sections:

1. A Reading Test focused on the assessment of students' comprehension and reasoning skills in relation to appropriately challenging prose passages (sometimes paired or associated with one or more informational graphics) across a range of content areas, as defined by the students' grade level.
2. A Writing and Language Test focused on the assessment of students' revising and editing skills in the context of extended prose passages (sometimes associated with one or more informational graphics) across a range of content areas, as defined by the students' grade level.

The optional Essay, which is only found on the SAT, is focused on the assessment of students' skills in developing a cogent and clear written analysis of a provided source text. The Essay offers scores that complement those from the other two English language arts/literacy assessments.

A number of key design elements strongly supported by evidence are interwoven throughout the Evidence-Based Reading and Writing portion and the Essay portion of the tests. These include:

- The use of a specified range of text complexity aligned to college and career readiness levels of reading
- An emphasis on source analysis and use of evidence
- The inclusion of data and informational graphics, which students must analyze in conjunction with text
- A focus on words in context and on word choice for rhetorical effect
- Attention to a core set of important English language conventions and to effective written expression
- The requirement that students work with texts across a wide range of disciplines

Several of these elements are Key Features of the SAT Reading Tests, as described in Chapter 3: Test Development Procedures and in extended discussions of test specifications (College Board, 2014). Other elements, while not explicitly Key Features, are nonetheless important factors and have been incorporated into the assessment in a significant manner, in keeping with what research shows to be most essential for college readiness.

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 SAT Suite Reading Test and the SAT Suite Writing and Language Test and the SAT Essay align the levels of text complexity represented in the tests' passages with the requirements of first-year college courses and workforce training programs. This alignment supports the emerging movement to close this preparedness gap by making text complexity a central part of the test design. Students taking the SAT Suite Reading Tests and SAT Suite Writing and Language Tests are asked to engage with passages selected, in part, to exhibit a range of text complexities up through and including levels comparable to those expected of students entering college and workforce training programs. Students taking the SAT Essay are asked to engage with a passage that is rich and challenging but not so difficult that high school juniors and seniors can't produce an effective written response to it. To ensure that texts on the SAT Suite are appropriately complex—challenging but not inaccessible to college- and career-ready test takers—test development staff make use of feedback from secondary and postsecondary subject-matter experts and test data on student performance, as well as quantitative and qualitative measures of text complexity. Considered together, the SAT Suite Reading Test, SAT Suite Writing and Language Test, and the SAT Essay measure whether students can read, improve, and analyze texts at levels of difficulty required of incoming postsecondary students.

Source Analysis and Evidence Use

Students' 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. National curriculum surveys conducted by the 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, 2006; Kim, Wiley, & Packman, 2012; Seburn, Frain, & Conley, 2013). Institutions such as Duke University, Cornell University, Texas A&M University, and the University of California, Berkeley, have devoted considerable resources to developing the skills of source analysis and evidence use in their students.

The SAT Suite Reading Tests, the SAT Suite Writing and Language Tests, and the SAT Essay support an emphasis on source analysis and evidence use throughout the Evidence-Based Reading and Writing portion and Essay portion of the tests. The SAT Reading Tests not only require students to answer questions based on what is stated and implied in texts (both passages and graphics) across a range of content areas, but also include a number of questions asking students to determine which portion of a text best supports the answer to a given question. The SAT Writing and Language Tests include questions asking students to develop, support, and refine claims and ideas in multiparagraph passages (some of which are associated with one or more graphics) and to add, revise, or delete information in accordance with rhetorical purpose and accuracy of content (as, for example, when students are asked to verify or improve a passage's explanation of a data table). In the SAT Essay, students are required to analyze a provided source text to determine how the author builds an argument to persuade an audience through the use of evidence, reasoning, and/or stylistic and persuasive devices (and potentially other aspects of the text identified by students themselves) and then to write a cogent and clear analysis supported by critical reasoning and evidence drawn from the source.

Analysis of Data in Graphics

The ability to understand and analyze quantitative information and ideas expressed graphically in tables, graphs, charts, and the like is an essential skill for college- and career-ready students. Friel, Curcio, and Bright (2001), for example, note that " . . . the use of visual displays of quantitative data is pervasive in our highly technological society" and observe that " . . . to be functionally literate, one needs the ability to read and understand statistical graphs and tables" (p. 124). Gal (2002) echoes these sentiments, writing that " . . . statistical literacy is a key ability expected of citizens in information-laden societies, and is often touted as an expected outcome of schooling and as a necessary component of adults' numeracy and literacy" (p. 1).

Attention to analysis of graphically displayed data is supported in part by incorporating data graphics into selected SAT Suite Reading Test and SAT Suite Writing and Language Test passages and questions. In the SAT Suite Reading Tests, students are expected to analyze and interpret data in tables, graphs, charts, and so on and to synthesize information and ideas presented graphically with those presented in a prose passage. In the SAT Suite Writing and Language Tests, students are asked to make particular choices about revising and editing prose passages in light of accompanying graphics. On this test, students may, for example, be asked to recognize and correct an error in a passage's interpretation of a table or to evaluate a graph's potential relevance to the topic of or claims in a passage. Coupled

with the graphics used in the SAT Suite Math Tests, the graphics in the SAT Suite Reading Test and the Writing and Language Test assess students' capacity to analyze quantitative data across a wide range of content areas.

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 almost double that frequency: 30.9 (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 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. (By contrast, Tier One words require little instruction for most students 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 Evidence-Based Reading and Writing and Essay portions of the tests. In the SAT Suite Reading Tests, students are called on to determine the meaning of vocabulary in context, with an emphasis on Tier Two words and phrases. In the Reading Tests, the Writing and Language Tests, and the SAT Essay, students are also presented with other vocabulary-related challenges, including analyzing word choice rhetorically; improving the precision, concision, and context appropriateness of expression; and (in the Essay) using language to convey their own ideas clearly and carefully.

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 written English and, more generally, making informed, thoughtful language choices. Knowledge of conventions includes learning and adhering to language "rules" set forth in textbooks, as well as knowledge of the conventions that lend precision and clarity to writing, aid comprehension, and facilitate academic success. Language conventions have been described in terms of grammatical choices as representing the relationships between writers and their world, as expressions of how writers attend to the words of others and position themselves in relation to others, and requiring cognitive skills at the level of idea development and at the sentence level (Micciche, 2004).

The SAT Suite tests support a thoughtful emphasis on language conventions and language use in several important ways. Effective language use and mastery of a core set of conventions linked with college and career readiness are two key elements of the Writing and Language Tests, which, among other aims, assess students' application of these skills

in the context of high-quality multiparagraph passages that must be revised and edited. The Reading Tests include questions that address students' capacity to analyze word choice rhetorically. The SAT Essay includes effective language use among the criteria for evaluating students' written analyses of source texts.

Disciplinary Literacy

Shanahan, Shanahan, and Misischia (2011) are prominent among those who have made the case in recent years that students' literacy development shouldn't 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 is 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). For example, instructors of entry-level health courses surveyed in California rated the knowledge of appropriate terminology in the healthcare setting as being very important for postsecondary success in the health sciences and medical technology career cluster (McGaughy et al., 2012).

The SAT Suite tests support an enhanced emphasis on disciplinary literacy through careful passage selection and question development. In the Reading Tests, the Writing and Language Tests, and the SAT Essay, students are expected to engage with and analyze appropriately challenging texts spanning numerous content areas, including U.S. and world literature, history/social studies, the humanities, science, and careers-related topics. Moreover, while questions on the Reading Test and the Writing and Language Test don't require students to have prior knowledge of specific topics in the content areas, these questions do, where possible and beneficial, reflect differences in the ways different disciplines approach literacy (e.g., Problems Grounded in Real-World Contexts, Analysis in Science and History; see Section 1.2). Reading 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. Reading questions relating to a science selection, on the other hand, might require students to delineate the experimental process described in a text, analyze research data (including data represented graphically), or determine which conclusion is best supported by a study's findings—skills generally not required to comprehend literary texts.

Evidentiary Foundation for the SAT Suite Math Tests

The overall aim of the Math 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 defined by their grade level.

As with Evidence-Based Reading and Writing, a number of key design elements strongly supported by evidence are interwoven through the Math area. Among these are:

- A focus on content that matters most for college and career readiness
- An emphasis on problem solving and data analysis
- The inclusion of both calculator and no calculator sections as well as attention to the use of a calculator as a tool

Several of these elements are key features of the Math Tests, as described in Section 3.1 and in extended discussions of test specifications (College Board, 2014). Other elements, while not necessarily key features are nonetheless important factors and have been incorporated into the assessment in a significant manner, in keeping with what research shows to be most essential for college readiness.

Focusing on Content That Matters Most

There is a major disconnect today in mathematics between the K–12 and higher education systems. In a recent 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% of high school teachers said yes, only 4% 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. As one survey noted,

> Because the postsecondary survey results indicate that a more challenging treatment of fundamental content knowledge and skills needed for credit-bearing college courses would better prepare students for postsecondary school and work, states would likely benefit from examining their state standards and, where necessary, reducing them to focus only on the knowledge and skills that research shows are essential to college and career readiness and postsecondary success (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" (Council of Chief State School Officers, 2013, p. 2).

One of the most important ways the SAT Suite Math Tests address the gap between postsecondary and K–12 expectations is through the assessment's concentrated focus on the content that matters most for postsecondary education. In a national survey published in 2011, Conley reinforced the conclusion that some content areas require much stronger emphasis than others. The distinctive importance of algebra is unmistakable from Conley's data. Other math domains have a more mixed profile, typically including more material that isn't as relevant to most postsecondary work and/or isn't a prerequisite for most postsecondary work. The data from this study directly support the content choices made in the Math Tests (Conley et al., 2011).

Problem Solving and Data Analysis

There is ample evidence that problem solving and data analysis—the ability 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. Quantitative literacy is part of participation in a democracy; it is important to employers, who need students who can use mathematics outside of the classroom; and it is 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).

A recent 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 SAT Suite Math 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, densities, and gradients). The Problem Solving and Data Analysis portions of the Math Tests contain multipart problems.

Calculator and No Calculator Portions

A calculator is a tool, and decisions about when and when not to use it involve a variety of considerations. The data are clear that postsecondary instructors expect students to be fluent in rational number arithmetic (ACT, Inc., 2007, 2009). Including the no calculator sections on the SAT Suite Math Tests helps assure postsecondary instructors that students who earn high scores on the SAT Suite tests don't lack the basic prerequisites. Questions in the calculator portion of the test are designed to probe students' ability to make wise choices between when to use the calculation and when not to use it. For some questions, the calculator lends efficiency; for others, the ability to make use of structure or to reason abstractly leads to the most rapid solution.

7.3 The Relationship Between the SAT and Other Similar Assessments

As part of determining that scores from the new SAT are valid for intended uses, the College Board used equipercentile methods to link scores from the new SAT with scores from the old SAT. AERA/APA/NCME Standard 5.18 states that "when linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly" (AERA/APA/NCME, 2014, p. 106). In keeping with this standard, the first part of this section discusses the concordance methods used to link the new SAT to the old SAT, as well as the new PSAT/NMSQT to the old PSAT/NMSQT, while the second part describes the vertical scaling methods used to link the new SAT to the new PSAT/NMSQT, PSAT 10, and PSAT 8/9.

How Does the New SAT Relate to the Old SAT? Background and Purpose of the Concordance Study

While redesigning the SAT, the College Board put forth significant efforts to examine score comparability and scale continuity issues. We also made mindful and deliberate efforts in communicating the changes between the old SAT and the new SAT to various stakeholders and the general public during the redesign process. The changes made to the old SAT are an effort to more accurately test a student's readiness for college and career. The new SAT differs from the old SAT in several key aspects, with the most notable differences being the content of the test itself. An in-depth discussion of how the new SAT differs from the old

SAT is found in Chapter 1: Overview. Comparisons of the features, scores, scales, and test lengths of the new SAT and the old SAT are also available as part of College Board's online SAT resources (College Board, 2015c).

Due to the nature and the scope of the differences between the new SAT and the old SAT, using the term "equating" isn't appropriate, as measures from the old SAT and the new SAT don't result in scores that can be made interchangeable. In order to supply a table that relates the scores from the two measures under these circumstances, a procedure known as concordance is performed (e.g., Dorans, 2000; Kolen & Brennan, 2004). Two concordance studies were conducted as part of the assessment redesign, an early preliminary study and a second final study. This section describes the methods and results from the final concordance study.

The concordance tables were prepared based on the internal College Board research for various intended uses including benchmark research, score reporting, and predictive validity. They are also used by external stakeholders including, but not limited to, high school counselors, higher education, and the National Collegiate Athletic Association (NCAA) for score reporting and other purposes. The SAT concordance tables are also used as concordance tables for the new and old scores of PSAT/NMSQT and PSAT 10.

Methods

Concordance Study Design and Methodology

Concordance data were collected using the new SAT and the October, November, and December 2015 administrations of the old SAT. Using a single group design, one sample of participants completed an intact form from both measures (e.g., old SAT and new SAT). Various methodologies for linking exist as possible options for selecting a concordance table solution [e.g., equipercentile, linear, with smoothing or without smoothing (Dorans, 2000; Kolen & Brennan, 2004)]. The ideal approach, which was used here, is to examine multiple solutions and select the most appropriate one. The criteria for deciding which concordance table to choose include coverage of the full area score scale (e.g., 200–800) and minimizing gaps between scores (e.g., scale scores skip from 680 to 700). Using the equipercentile scaling method with cubic spline postsmoothing, the percentile ranks of the scale score points from the new SAT and the old SAT are linked to one another to determine which scores correspond from the two forms. For this study, 11th and 12th graders completed both the new SAT and the old SAT and then analyses were conducted to produce concordance tables linking the scores across these two forms.

Instruments

The two instruments used in the concordance study were the old SAT and the new SAT. Specifically, the examinees who completed the October 2015, November 2015, or December 2015 administration of the old SAT were recruited to take the new SAT on December 9, 2015. This was a single group design where students who took the new SAT on December 9 also had a previous old SAT score from October, November, or December 2015. If students completed more than one of the three old SAT administrations, only their scores from the most recent administrations were kept for the purpose of this study. The new SAT has five types of scores: (1) the total score, (2) two section scores, (3) two cross-test scores, (4) three test scores, and (5) seven subscores (for the specific concordances produced, see Table A-7.1 in Appendix 7: Validity). Concordances weren't created for the cross-test scores or the subscores. Survey questions that ask about motivation level and the necessary demographics were administered as part of the answer sheets.

Participants

The sampling plan targeted 50% 11th graders and 50% 12th graders. A high-level sampling plan with target specifications for region, location, public vs. private, grade level, and percent underrepresented minority students was also created and intended to represent a typical SAT cohort. All schools had to have sufficient training on administering secure assessments. Students with disabilities were allowed to participate in the study as well, but weren't included in the analyses.

The new SAT administration in December 2015 was conducted as an operational administration, and the scores students received from this administration were reportable. The target population was a typical SAT cohort, defined as the average of the 2011–2014 old SAT cohorts. The cleaned unweighted data set was composed of N = 8,677 examinees (see Table A-7.2 in Appendix 7: Validity for the score correlations and Tables A-7.3 through A-7.9, also in this appendix, for more detailed descriptions of the unweighted data). A weighted version of this data was created to better represent the target population. Specifically, the 8,677 examinees' case weights were obtained, which resulted in a weighted data set that approximated recent SAT cohorts in terms of percentages of demographic variables such as grade, gender, ethnicity, region, best language, and first language, as well as the means and standard deviations of the old SAT scale scores (see Table A-7.2 in Appendix 7: Validity for the score correlations and Tables A-7.10 through A-7.16, also in this appendix, for more detailed descriptions of the weighted data).

Analysis, Methodology, and Results

The Linking with Equivalent Groups or Single Group Design (LEGS) software written by the Center for Advanced Studies in Measurement and Assessment (CASMA) (Brennan, 2004) was used in the concordance analysis, with the specific analytic steps described below.

Step 1: Data Cleaning and Screening

Four data files were available for the analyses: three for the old SAT (one from each administration) and one for the new SAT. After matching students' old and new SAT scores using "person_id," and using only the most recent old SAT score per student, a sample of 8,776 students was obtained. The final sample of N = 8,677 was obtained after applying the additional data screenings as listed below:

- Removed SSD answer sheets (N = 64)[1]
- Removed students indicated as having irregularities (N = 23)
- Removed 10th graders (N = 9)
- Removed examinees who skipped the entire Reading, Math, or Writing tests of the new SAT (i.e., operational screening rules; N = 3)

Step 2: Determine the Representativeness of the Sample and Weight the Sample Appropriately

At this point, the cleaned sample was ready to be analyzed for the concordance. The next step was to determine if this sample closely approximated the target population and, if not, to determine a weighted sample that would be a better approximation of the target population. As mentioned previously, the uncleaned unweighted data set was weighted by the variables of interest that were deemed practically meaningful. The means and standard deviations

[1] In this context, "SSD answer sheets" are defined as the answers sheets from test takers who were provided testing accommodations by the Services for Students with Disabilities office.

of the old SAT scores were examined before and after the sample data were weighted. As expected, the weighted data set was a closer approximation of the target population and was used as input data for all correlation and concordance analyses.

Step 3: Correlation Analyses

Table A-7.2 in Appendix 7: Validity shows the correlations between the score pairs of interest for this concordance study. The lower diagonal is based on the unweighted data, and the upper diagonal is based on the weighted data (which were used as inputs into the concordance analyses). Any value set in boldface shows a correlation between two score pairs that were used to create a concordance table.

Step 4: LEGS Analyses and the Resulted Concordance Solution for SAT

All concordance analyses were done using the Linking with Equivalent Groups or Single Groups (LEGS) software program. For each score pair shown in Table A-7.1, two versions of the concordance tables are provided:

1. The "old SAT to new SAT" concordance table
2. The "new SAT to old SAT" concordance table

The unidirectional "old SAT to new SAT" and "new SAT to old SAT" tables were obtained by conducting two separate LEGS runs. The "old SAT to new SAT" tables include every possible old SAT score and provide a concorded new SAT score for each. The "new SAT to old SAT" tables include every possible new SAT score and provide a concorded old SAT score for each. Note that while each of these tables contains all possible values for the *from* score (e.g., the old SAT score in the old SAT to new SAT table) because of differences in number of score points on the scales compared and the resulting gaps and many-to-one concordances, each table doesn't contain all score points for the concorded *to* score (e.g., the new SAT score in the old SAT to new SAT table). Cubic spline postsmoothing was used for all concordance tables. All the SAT concordance tables and their plots are provided in Tables A-7.17 to A-7.28 and Figures A-7.1 to A-7.12 in Appendix 7: Validity. Several criteria were considered when determining the recommended final concordance table, specifically:

- Top of the scale must be represented.
- Bottom of the scale must be represented.
- Many-to-one conversions should be minimized.
- Score gaps should be minimized.

As implemented in LEGS, the cubic spline postsmoothing method is replaced with linear interpolations for the lowest and highest scores with sparse data (usually 0.5%, meaning that linear interpolations are applied for *from* scores with percentile ranks less than 0.5 and above 100-0.5 = 99.5 [Kolen & Brennan, 2004]). The extent of interpolation was manipulated and selected to achieve a minimal number of many-to-one conversions in each concordance table.

PSAT/NMSQT and PSAT 10 Concordance Tables

The PSAT/NMSQT (administered in the fall) and the PSAT 10 (administered in the spring) use the same test forms, with scores linked with the old PSAT/NMSQT by the same set of concordance tables. Due to the vertical scales that were established between the new SAT and the PSAT/NMSQT and PSAT 10 (described in detail in the next section),

the SAT concordance tables described earlier in this chapter were also used as the final PSAT/NMSQT and PSAT 10 concordance tables. Just as with the SAT concordance table, there are two versions of each PSAT/NMSQT concordance table provided:

1. The "old PSAT/NMSQT to new PSAT/NMSQT and PSAT 10" concordance table.
2. The "new PSAT/NMSQT and PSAT 10 to old PSAT/NMSQT" concordance table

The only difference between these tables and the SAT concordance tables is that these tables have scale score ranges adjusted to reflect the old PSAT/NMSQT (e.g., a 400 on Math for the old SAT is a 40 for Math on the old PSAT/NMSQT). Also, the lowest value of these scale score ranges for the new PSAT/NMSQT and PSAT 10 is lower than the new SAT, so those scores that are part of the new PSAT/NMSQT and PSAT 10 vertical scale but not the new SAT vertical scale have interpolated values associated with them in the "new PSAT/NMSQT and PSAT 10 to old PSAT/NMSQT" concordance tables. All of the PSAT/NMSQT and PSAT 10 concordance tables are provided in Tables A-7.29 –A-7.40 in Appendix 7: Validity.

ACT–SAT Concordance Tables

By merging the concordance tables that relate the old SAT and the new SAT with the existing ACT–SAT concordance tables, an additional set of concordance tables that didn't involve any new data collection was produced (College Board, 2009). Specifically, the results of the final concordance tables from this December 2015 concordance study were leveraged to connect the existing ACT–SAT concordance tables to the new SAT scale scores. In other words, these concordance tables are derived concordances between the ACT and the new SAT. There are two sets of the existing concordance tables, one for the ACT composite and the old SAT total score (defined as Critical Reading plus Math), and a second for the ACT Writing (pre-September 2015, before the recent ACT Writing redesign) and the old SAT Writing. These new derived concordance tables are extensions of the existing tables with the intention of adding the new SAT. However, only single score points are included (whereas score ranges were in the existing concordance tables for the old SAT). Also, the newly derived concordance tables are shown in two directions—ACT to new SAT and new SAT to ACT. There are four tables: (1) ACT Composite to new SAT Total (2) ACT Writing from pre-September 2015 to new SAT Writing and Language, (3) New SAT total to ACT Composite and (4) New SAT Writing and Language test to ACT Writing from pre-September 2015. These tables are shown in Tables A-7.41 to A-7.44 in Appendix 7. For lower score points on these tables, there wasn't enough valid data to produce a valid concordance between the new SAT and ACT. A new concordance study for the ACT and new SAT will be conducted in the future, using actual students' data on the ACT and the new SAT. ACT and the College Board are committed to working together to create a set of concordance tables between the ACT and the new SAT, expected by 2018. In order to provide the most accurate results, the concordance tables will be based on an analysis of scores from tests taken approximately during the first full year of administration of the new SAT, March 2016 through June 2017 and a similar full year of ACT administrations. In the interim, to assist users during this transition phase to the new SAT scale, these derived concordance tables were created.

SAT Concordance Table Implications and Cautions

The concordances should be interpreted with the following cautions:

1. Concordance tables are sample dependent. The sample that was used to create these tables took the old SAT in October, November, or December 2015, followed by the new SAT in a special administration on December 9, 2015. For this sample, the old SAT

scores were the test takers' most recent from October, November, or December 2015 and the new SAT scores were their first time taking this test. However, we weighted the sample according to demographic characteristics of recent SAT cohorts, so the sample used for the concordance study represents this target population, although the data were collected in the administration-specific samples.

2. Concordances are needed for students who take either the old SAT or the new SAT. An order effect was unavoidable in the data collection design, and thus, the concordance data sets were assembled from students who took the old SAT followed by the new SAT. It wasn't possible to counterbalance the order of administrations, or more ideally have an equivalent groups design with each student taking only one of the tests being concorded. Thus, interpretation of these tables should be done cautiously, keeping in mind the order effects resulting from the data collection limitation.
3. The concordance tables were developed using the total sample, which includes various demographic groups, and they need to be evaluated for invariance with respect to subgroups. The College Board has conducted preliminary invariance evaluation for concordance tables established for the SAT Suite of Assessments. Additional invariance evaluations will be conducted to assess the accuracy of these concordances for examinees in several demographic subgroups.

How Does the SAT Relate to PSAT-Related Assessments?

The new PSAT/NMSQT, PSAT 10, and PSAT 8/9 scales were established to support descriptions of student growth on the new SAT scales (Chapter 1: Overview and Section 6.1). The vertical scaling process utilized a scaling test design (Kolen & Brennan, 2014) where examinees took a complete test form in PSAT/NMSQT, PSAT 10, or PSAT 8/9 along with one of five randomly assigned scaling tests developed to represent the content domain and statistical characteristics of the combined SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9. The vertical scales were based on directly linking the number correct scores of a new PSAT/NMSQT and PSAT 10 base form and a new PSAT 8/9 base form to the new SAT scales for the Math Section, the Reading and Writing and Language Tests, and the Analysis in Science and Analysis in History/Social Studies Cross-Tests.[2] Because PSAT/NMSQT and PSAT 10 administer the same test forms, the same vertical scaling results were obtained for PSAT/NMSQT and for PSAT 10.

The vertical scaling process had the following steps. These steps will be described in more detail in a forthcoming monograph documenting the scaling procedures (College Board, 2016b).

1. The scaling study data were obtained for the examinees who took the base form used in establishing the new SAT scales (Section 6.1). These test takers had scores on all sections of the new SAT test (see Figure 5.1 in Section 5.1), and scores on an additional, randomly administered scaling test was developed to be representative of the content on Math, Reading, Writing and Language, Analysis in Science or Analysis in History/Social Studies across the new SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9 assessments.
2. Other data were obtained for examinees who took a base form for the new PSAT/NMSQT and PSAT 10, and for examinees who took a base form for the new PSAT 8/9. These test takers took one of the five additional scaling tests administered with the SAT scaling

[2] Subscores for PSAT/NMSQT, PSAT 10, and PSAT 8/9 were not vertically scaled. Their scales were established using processes described in Section 6.1.

study examinees obtained in Step 1 (i.e., representative of the content on Math, Reading, Writing and Language, Analysis in Science, or Analysis in History/Social Studies across the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9 assessments).

3. The SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9 test taker data were screened to identify plausibly motivated test takers based on the percentages of test questions completed and their responses to a survey question on their motivation.
4. The test taker data were screened to obtain grade-level participation for the target populations of interest (i.e., 11th graders for SAT, 10th graders for PSAT/NMSQT and PSAT 10, and 9th graders for PSAT 8/9).
5. The screened SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9 test taker samples were weighted to approximate the same set of nationally representative demographics described in recent NCES publications. Weighting was used to make student proportions more similar to national proportions of private and public school students, geographic regions, urbanicities, gender, ethnicities, and college intention.
6. The number correct scores of the new PSAT/NMSQT and PSAT 10, and the new PSAT 8/9 forms were linked to those of the new SAT form through the scaling test using the chained equipercentile method. This process was implemented after presmoothing the test and scaling test distributions with log-linear models, and was completed separately for Math, Reading, Writing and Language, Analysis in Science, and Analysis in History/Social Studies.
7. SAT scale score conversions for the PSAT/NMSQT, PSAT 10, and PSAT 8/9 scores were obtained for the conversions from Step 6 by applying interpolations to the SAT scale score conversion tables (Section 6.1).
8. Rounding and score truncations were applied to obtain scale scores with the desired ranges for the PSAT/NMSQT and PSAT 10 (160–760 for section scores and 8–38 for test/cross-test scores) and for PSAT 8/9 (120–720 for section scores and 6–36 for test/cross-test scores).

By equating additional new and nonscaled PSAT/NMSQT, PSAT 10, and PSAT 8/9 forms to the form used to establish the vertical scale, the established vertical scales are preserved (see Section 6.2).[3]

7.4 The Relationship Between the SAT and College Outcomes

Now that we've discussed the relationship between scores from the new SAT and scores from the old SAT, it is time to examine the relationship of those new SAT scores with key college outcomes.

Background Information on the Pilot Predictive Validity Study

An important use of the SAT is for college admission decisions. Therefore, it's critical to examine and document the relationship between SAT scores and college performance. Because operational new SAT scores (as well as the following year's college grades) aren't currently available for students, we can't conduct a traditional validity study. Instead, a pilot

[3] The standardized mean differences (SMD) in Tables A-7.4 to A-7.9 and A-7.11 to A-7.16 were calculated differently from those described in Section 6.5 and in Appendix 6. For the concordance results, the SMD values were computed as the difference in means of a Subgroup and the Total group divided by the standard deviation for the Total group.

study of the predictive validity of new SAT scores with first-year college grades was conducted. The results of this pilot study, which will be described in this chapter, show that the new SAT is as predictive of college success as the old SAT; that new SAT scores improve the ability to predict college performance above using just high school grade point average (HSGPA); and that there is a strong, positive relationship between new SAT scores on various sections of the assessment and the grades earned in matching college course domains. This suggests that the new SAT is sensitive to instruction in English language arts, math, science, and history/social studies. A more complete description of this pilot study can be found in Shaw et al. (2016).

Methodology

Study Design

A typical operational admission validity study would use students' recorded SAT scores and their reported first-year grade point average (FYGPA) to examine the statistical association between the two. Because this was a pilot study and not an operational validity study, it was necessary to first administer a pilot form of the new SAT to students who had just begun their first year of college. We then followed those students through this first year and collected their grades and FYGPA as the outcome for analyses. In order to do this, the College Board partnered with four-year institutions in the U.S. to administer the test and then collect student grades.

Participants

Institutional Sample. The goal for this study was to recruit 10–15 diverse four-year institutions for participation so that a sample of their students could participate in a campus administration of the new SAT. To design a sampling plan, we first outlined the population of four-year institutions from the College Board's *Annual Survey of Colleges* (ASC) from 2012, which collects information from colleges, universities, vocational/technical schools, and graduate schools that is of interest to potential applicants. The population of four-year institutions from which we sampled was specified as follows:

- Located within the United States
- Accredited by at least one accrediting agency
- Has at least 200 enrolled degree-seeking, first-year students who sent SAT scores to the institution
- Uses test information to make admission decisions
- Is either public or private (but not private, for-profit)
- Is a bachelor's degree–granting institution

Based on these criteria, the number of total eligible institutions from which to sample was 699. Institutions were then stratified by region, admission selectivity, institution size, and institution control (public or private) to determine sample targets. The desired sample of institutions was then developed to best reflect the population while also aiding in the study administration (e.g., larger institutions would have a more likely chance of recruiting students to participate in the study). The recruitment of institutions was facilitated by regional College Board staff who are closely connected to the colleges and universities. As the requirements for study participation were too burdensome for some institutions, similar institutions were identified as backup institutions in order to maintain as diverse and representative a sample as possible when selecting 10–15 institutions out of 699.

Table 7.1 provides information on the sample of institutions in the study. For a deeper analysis of the institutional sample and comparisons to the population and other samples, see Shaw et al. (2016).

Table 7.1: Institutional Sample Characteristics

		Pilot Study Sample (n_i = 15)	
		n_i	%
U.S. Region	Midwest	1	7
	Mid-Atlantic	2	13
	New England	2	13
	South	5	33
	Southwest	3	20
	West	2	13
Control	Public	10	67
	Private	5	33
Admittance Rate	Under 50%	6	40
	50%–75%	6	40
	Over 75%	3	20
Undergraduate Enrollment	Small	0	0
	Medium	5	33
	Large	2	13
	Very Large	8	53

Note: Percentages may not sum to 100 due to rounding. Undergraduate enrollment was categorized as follows: small—750–1,999; medium—2,000–7,499; large—7,500–14,999; and very large—15,000 or more.

Student Sample. Participating institutions were charged with recruiting as representative a sample of first-year students at their institution as possible. Students also had to have previously taken the SAT (in the 2014 college-bound seniors cohort) so that comparisons between their operational SAT scores and their new SAT scores could be made. This comparison was primarily used to identify students with particularly low motivation during the pilot test administration.

There were 2,182 students who participated in the new SAT test administration across the 15 institutions in the study. Thirty-two students were dropped from the sample because they: (1) didn't have an SAT score in the 2014 college-bound seniors cohort (n = 21), (2) didn't have an SAT score on record at all (n = 5), (3) were not first-time freshmen (n = 1), or (4) couldn't be matched from their test administration data to the College Board database (n = 5).

Among the 2,150 students who remained in the sample, additional filtering needed to take place to ensure that all students had the study variables of interest. There were 61 students who either didn't have an HSGPA on record (n = 57) or didn't have an FYGPA (n = 4); these students were removed from the study sample.

For the 2,089 remaining students, it was important to examine concerns with low student motivation, as this test was administered as part of a study as opposed to in a high-stakes condition. First, operational old SAT scores on record for students in the sample were concorded to new SAT scores using a concordance table linking scores on both tests (see Section 7.3). The difference of the actual new SAT score that the student took for the study from the concorded new SAT score was calculated and this difference value was then standardized for each student (the student's difference score minus the mean difference score, divided by the standard deviation of the difference score). This was done for the Evidence-Based Reading and Writing (ERW) section and the Math section. Standardized score differences that were greater than ±2 in either section were flagged. Another flag for low effort was created for students responding with "Disagree" or "Strongly Disagree" to the following item on the new SAT answer sheet, "I plan to put forth my best effort during the test today." The researchers determined that those students with an ERW score difference flag and a Math score difference flag should be dropped from the study. Also, those with either an ERW score difference flag (but no Math score difference flag) and a low effort flag were dropped, as well as students with a Math score difference flag (but no ERW score difference flag) and a low effort flag. Thirty-nine students were removed from the study based on these analyses. Therefore, the final sample included 2,050 students. See Table 7.2 for the characteristics of the student study sample. For a deeper analysis of the student sample and comparisons to the population and other samples, see Shaw et al. (2016).

Table 7.2: Student Sample Characteristics

		Pilot Study Sample (n_s = 2,050)	
		n_s	%
Gender	Male	743	36
	Female	1,306	64
Race/Ethnicity	African American	262	13
	American Indian	9	0
	Asian	403	20
	Hispanic	358	17
	Other	64	3
	White	949	46
	Not Stated	5	0
Best Language	English Only	1,653	81
	English and Another	361	18
	Another Language	26	1
	Not Stated	10	0

Table 7.2 continued on next page

Table 7.2 continued from previous page

		Pilot Study Sample (n_s = 2,050)	
		n_s	%
Parental Income	< $40,000	284	14
	$40,000–$80,000	309	15
	$80,000–$120,000	286	14
	$120,000–$160,000	126	6
	$160,000–$200,000	74	4
	> $200,000	102	5
	Not Stated	869	42
Highest Parental Education Level	No High School Diploma	86	4
	High School Diploma	420	20
	Associate Degree	121	6
	Bachelor's Degree	697	34
	Graduate Degree	692	34
	Not Stated	34	2

Note: Percentages may not sum to 100 due to rounding. One student in the pilot study sample didn't indicate gender.

Measures

New SAT Scores. New SAT scores were obtained from the special administrations of a pilot form of the new SAT in fall 2014 for this study. This includes the following scores:

Two section scores (200–800 scale)

Evidence-Based Reading and Writing (not including SAT Essay)—increments of 10

Math—increments of 10

Three test scores (10–40 scale)

Reading—increments of 1

Writing and Language (not including SAT Essay)—increments of 1

Math—increments of 0.5

Two cross-test scores (10–40 scale)

Analysis in Science—increments of 1

Analysis in History/Social Studies—increments of 1

New SAT Pilot Study Questionnaire Responses. Self-reported responses to questions on test day informed this study design, including questions related to motivation and effort, as well as student information, allowing researchers to match data from the pilot study to student's operational SAT scores on record.

SAT Questionnaire Responses. Self-reported gender, race/ethnicity, best language, parental education level, and parental income level were obtained from the SAT Questionnaire that students complete during registration for the operational SAT.

High School Grad Point Average (GPA). Self-reported HSGPA was obtained from the SAT Questionnaire when students had taken the operational SAT and is constructed on a 12-point interval scale, ranging from 0.00 (F) to 4.33 (A+).

College Grades. FYGPA and grades in all courses in the first year of college were obtained from the participating institutions. All courses were coded for content area so that analyses could be conducted on course-specific grade point averages. Course-specific grade point averages were calculated by student—across all relevant course grades received in a particular area during the first semester of college (excluding remedial coursework). For example, if a student took only one mathematics course in their first semester, then their average course grade in mathematics is based on the grade earned in that one course. If they took three mathematics courses, the average course grade is based on the average of the three course grades earned (taking into account the grades earned and the number of credits associated with each grade).

Analysis

The focus of this study is on providing validity evidence for the use of new SAT scores for college admission. Therefore, analyses were primarily correlational in nature and also graphical, depicting the relationships between the test scores and criteria of interest.

Correlational analyses were conducted to examine the strength of the relationship between the predictors of interest in the study (SAT scores and HSGPA) with FYGPA or college course grades. A correlation represents the extent to which two variables are linearly related and is on a scale of -1 to +1, where +1 is a perfect positive linear association and -1 is a perfect negative linear association. It is also helpful to think of a correlation as the extent to which a scatterplot of the relationship between two variables (e.g., SAT scores and FYGPA) fits a straight line (Miles & Shevlin, 2001).

Perfect linear associations essentially do not exist in applied social science research, so to contextualize the strength of correlation coefficients it is most helpful to either compare correlation coefficients to other correlations representing familiar or similar relationships (Meyer et al., 2001) or refer to a rule of thumb offered by Cohen (1988). Cohen's heuristic provides a quick way to evaluate the meaningfulness of an association or effect. Correlations that have an absolute value of approximately .1 are considered "small," correlations that have an absolute value of approximately .3 are considered "medium," and correlations that have an absolute value of .5 or greater are considered "large." Note that correlation coefficients (corrected for restriction of range) representing relationships between admission test scores and performance in college or graduate school tend to be in the .40s and .50s (Kuncel & Hezlett, 2007; Mattern & Patterson, 2014).

Bivariate and multiple correlations in this study were calculated; then these resulting correlation coefficients were also corrected for range restriction (both raw and corrected correlations are reported in this study). Admission validity research typically employs a correction for restriction of range because the variability of a set of predictors (e.g., SAT scores and HSGPA) is reduced due to direct or indirect selection on all or a subset of predictors. By definition, the narrowing of a score range by selection results in an underestimation of the true relationship between the predictor(s) and criterion (e.g., FYGPA). Mattern, Kobrin, Patterson, Shaw, and Camara (2009) noted that because applicants with higher HSGPAs or SAT scores are more likely to be admitted, the range of HSGPAs and SAT

scores is restricted compared to the range for the full applicant pool with those measures available. This study used the Pearson-Lawley multivariate correction for restriction of range (Gulliksen, 1950; Lawley, 1943; Pearson, 1902) with the 2014 college-bound seniors cohort to develop the unrestricted population covariance matrix for the correction.

Separate restriction-of-range corrected bivariate correlation matrices were computed for each participating institution instead of across all participating institutions. These separate matrices were then used to calculate the multiple correlations between the predictors and criterion as well as the average bivariate and multiple correlations, which were weighted by institution sample size.

Of particular interest in this study were the relationships between the different SAT scores (as well as all SAT section scores together) and FYGPA, as well as the incremental or additional validity that the SAT adds to the prediction of FYGPA above HSGPA. By examining the difference between the HSGPA–FYGPA correlation and the multiple correlation of SAT and HSGPA together with FYGPA, the incremental validity of the SAT is estimated. When possible and appropriate, relationships between SAT scores and criteria of interest are also presented graphically to more clearly show trends and relationships.

Results

First-Year Grade Point Average

Descriptive statistics for the academic variables were calculated for the student sample. Table 7.3 shows that this is an academically strong sample with a mean HSGPA of 3.85 and mean SAT section scores of 621 (SD = 100) for ERW and 634 for Math (SD = 113). The mean FYGPA for the study sample was 3.30 (SD = 0.60). For reference, in the 2012 SAT Validity Study sample (Beard & Marini, 2015), the mean HSGPA was 3.62 (SD = 0.50) and the mean FYGPA was 3.02 (SD = 0.72).

Table 7.3: Descriptive Statistics for Study Variables

	Mean	SD	Min	Max
HSGPA	3.85	0.43	1.67	4.33
SAT Total Score	1254	201	570	1600
SAT Evidenced-Based Reading and Writing Section	621	100	290	800
Reading Test	31	5.3	15	40
Writing and Language Test	31	5.2	11	40
SAT Math Section	634	113	230	800
Math Test	32	5.7	11.5	40
SAT Analysis in Science	31	5.2	13	40
SAT Analysis in History/Social Studies	31	5.3	12	40
FYGPA	3.30	0.60	0.00	4.17

Note: n = 2,050.

Table 7.4 shows the intercorrelation matrix for the primary predictors of interest for this study. HSGPA is correlated with both SAT sections (.50 for ERW and .49 for Math), indicating that there is a strong relationship between the two measures but that they aren't precisely measuring the same thing.

Table 7.4: Corrected (Raw) Correlation Matrix of Redesigned SAT Sections and HSGPA

	HSGPA	SAT ERW	SAT Math
HSGPA			
SAT ERW	.50 (.23)		
SAT Math	.49 (.23)	.77 (.60)	

Note: n = 2,050. Restriction of range-corrected correlations are presented. The raw correlations are shown in parentheses.

Table 7.5 depicts the corrected and raw correlations of the study predictors with the primary outcome of interest in this study, FYGPA. Confidence intervals for the corrected correlations are also presented to display the range of correlations within which we would expect the population correlation to be found with 95% confidence. Based on Cohen's (1988) rules of thumb for interpreting correlation coefficients presented earlier, you can see that the correlations between HSGPA and SAT scores with FYGPA are large, with the strongest relationship represented by the multiple correlation of both HSGPA and SAT together (r = .58). In this sample, the multiple correlations of the SAT ERW and Math sections together with FYGPA is .53, while the correlation between HSGPA alone and FYGPA is .48.

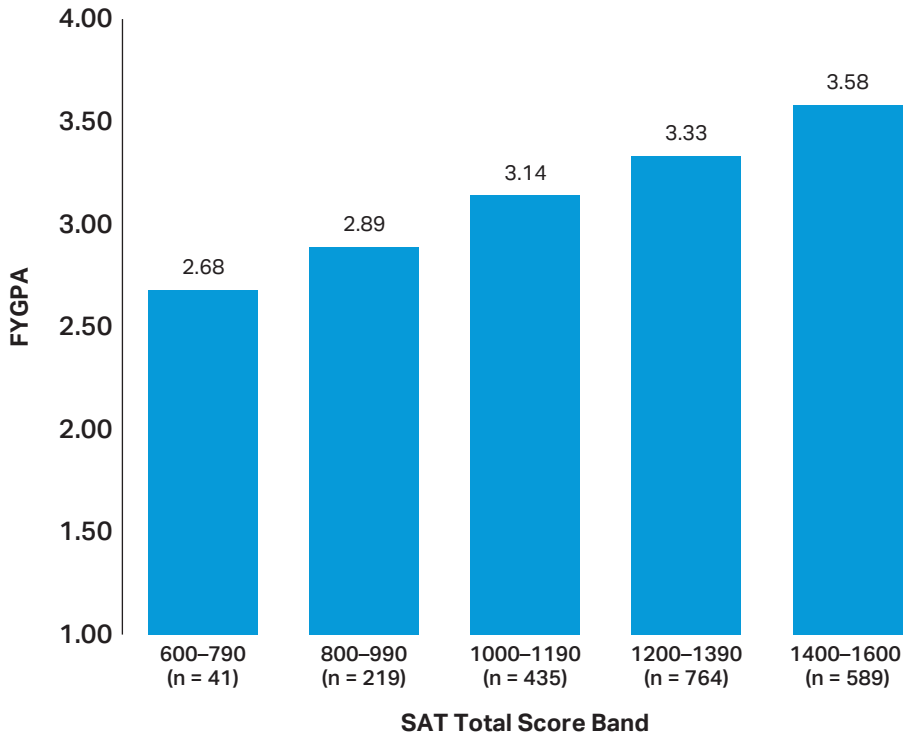
Table 7.5: Correlations of Predictors with FYGPA

Predictors	Correlations	95% Confidence Interval for Corrected Correlations
HSGPA	.48 (.27)	[.45, .51]
SAT ERW Section Score	.51 (.33)	[.48, .54]
SAT Math Section score	.49 (.30)	[.46, .52]
SAT ERW, SAT Math	.53 (.35)	[.50, .56]
HSGPA, SAT ERW, SAT Math	.58 (.40)	[.55, .60]

Note: n = 2,050. Pooled within institution, restriction of range-corrected correlations are presented. The raw correlations are shown in parentheses. The confidence intervals for bivariate correlations were calculated using the Fisher's Z transformation. Confidence intervals for the multiple correlations were calculated using the MBESS package in R.

To more easily understand what a correlation of .53 represents, you can examine Figure 7.1, which shows the average FYGPA that students earn by SAT total score band. In this figure, it is clear that as the SAT score band increases, there are corresponding increases in mean FYGPA. For example, those students with an SAT score between 800 and 990 earned, on average, an FYGPA of 2.89, while those students with an SAT score between 1400 and 1600 earned, on average, an FYGPA of 3.58.

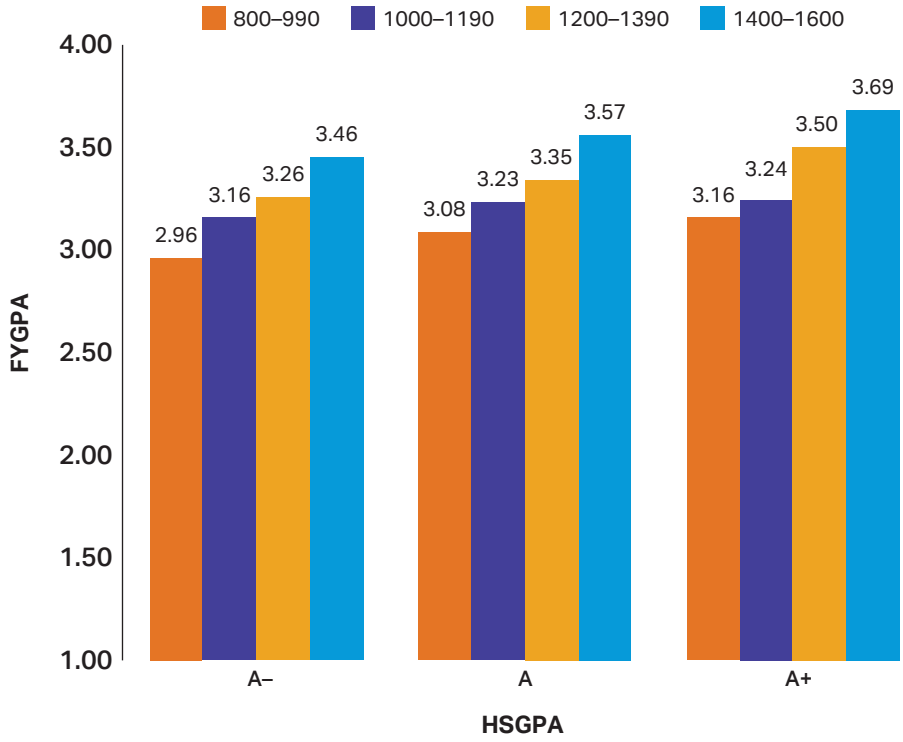
Figure 7.1: Mean FYGPA by SAT total score band



Note: Results based on fewer than 15 students aren't reported (e.g., score band 400–590, n = 2).

Note that the incremental validity added by the SAT above HSGPA is .10 (calculated from the difference between the multiple correlations of SAT and HSGPA with FYPGA of .58 and the HSGPA correlation with FYGPA of .48). To more easily understand what this incremental validity of .10 represents, Figure 7.2 graphically depicts the mean FYGPA by SAT total score band, after controlling for HSGPA by grouping students into the same HSGPA categories (among all students who received an A). In this figure, you can see that even within students grouped by the same HSGPAs of A-, A, or A+ (representing 84% of the study sample) there is a clear positive relationship between the SAT score bands and mean FYGPA. If there were no added value to having the SAT in addition to HSPGA in understanding students' FYGPAs, you would expect that all SAT score bands within HSGPA would have the same mean FYGPA value. Instead, for example, you can see that among those students with an 'A' HSGPA, those in the SAT total score band of 800–990 have a mean FYGPA of 3.08, while those same 'A' students in the SAT total score band of 1400–1600 have a mean FYGPA of 3.57.

Figure 7.2: Mean FYGPA by SAT total score band, controlling for HSGPA



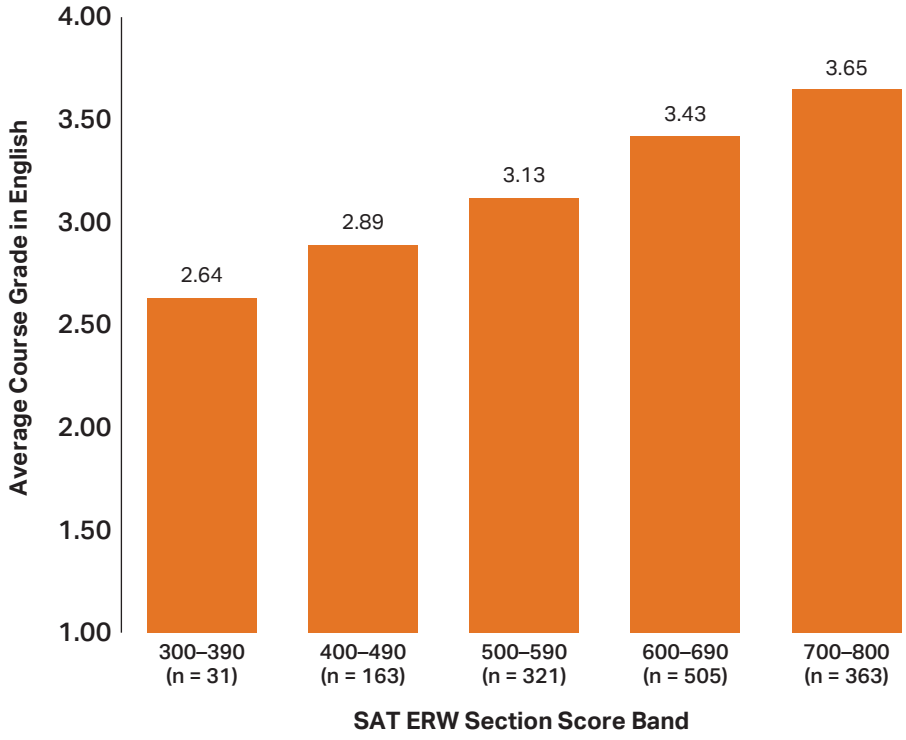
Note: HSGPA ranges are defined as follows: A- = 3.67 (or 90–92), A = 4.00 (or 93–96), and A+ = 4.33 (or 97–100). Results based on fewer than 15 students aren't reported; not reported are all students in the 400–590 and 600–790 SAT total score bands.

Course-Specific Grade Point Average

In addition to understanding the relationships between SAT scores and FYGPA based on correlational analysis, we explored the relationships between SAT section and cross-test scores with average first-semester course grades in the matching domain using graphical representations. All student coursework data in this study were coded for their content area focus as well as whether or not they were remedial courses. Remedial coursework wasn't included in this analysis.

Content experts, assessment developers, and researchers then worked to match the appropriate coursework codes with the matching SAT scores so that the relationship between the scores and college performance in the matching content area could be examined. Figure 7.3 shows the relationship between SAT ERW scores and average first-semester, credit-bearing college course grades in reading- and writing-intensive courses, including history, literature (not composition), social science, and writing courses. This graph depicts a clear positive relationship between SAT ERW scores and grades in matching college courses. For example, those students with an SAT ERW score of 400–490 have an average matching college course grade of 2.89, whereas those students with an SAT ERW score of 700–800 have an average matching college course grade of 3.65.

Figure 7.3: Relationship between SAT Evidence-Based Reading and Writing scores and course grades in the same domain

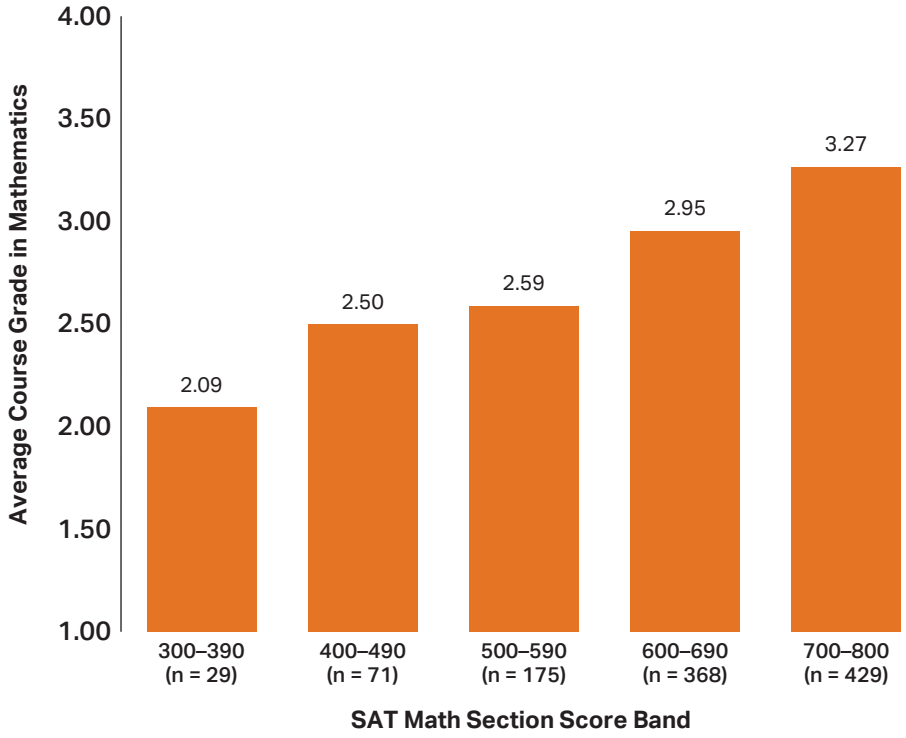


Note: Results based on fewer than 15 students aren't reported (e.g., score band 200–290, n = 1). Average English course grade includes first-semester courses that are reading and writing intensive (excluding foreign and classical languages).

Figure 7.4 shows the relationship between SAT Math scores and average first-semester, credit-bearing college course grades in algebra, precalculus, calculus, and statistics. This graph depicts a clear positive relationship between SAT Math scores and grades in matching college courses. For example, those students with an SAT Math score of 400–490 have an average matching college course grade of 2.50, whereas those students with an SAT Math score of 700–800 have an average matching college course grade of 3.27.

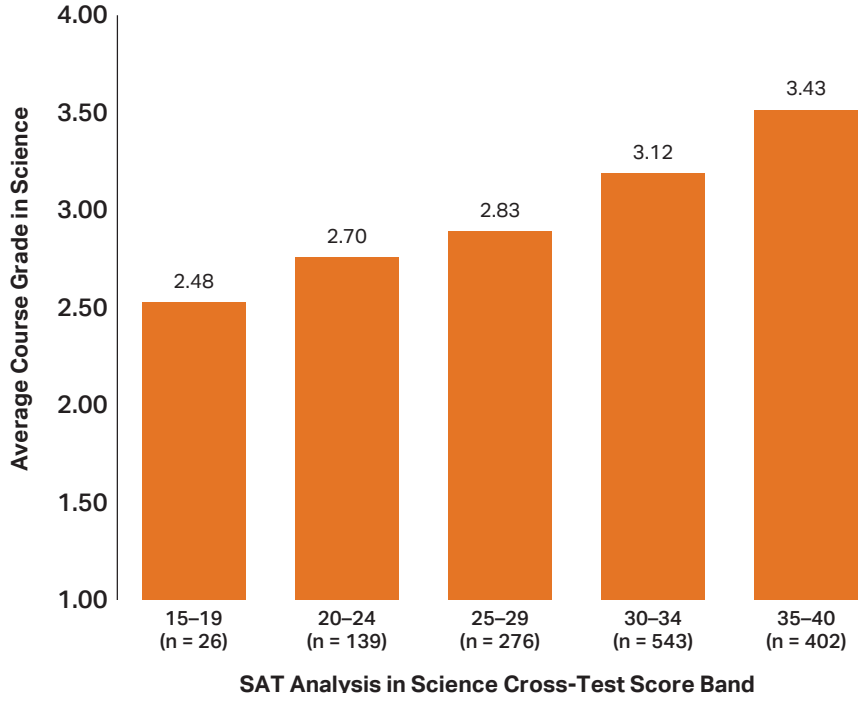
Figure 7.5 shows the relationship between SAT Analysis in Science cross-test scores and average first-semester, credit-bearing college course grades in science, including natural sciences, health sciences, and engineering. This graph depicts a clear positive relationship between SAT Analysis in Science cross-test scores and grades in matching college courses. For example, those students with an SAT Analysis in Science cross-test score of 20–24 have an average matching college course grade of 2.70, whereas those students with an SAT Analysis in Science cross-test score of 35–40 have an average matching college course grade of 3.43.

Figure 7.4: Relationship between SAT Math Section scores and course grades in the same domain



Note: Results based on fewer than 15 students aren't reported (e.g., score band 200–290, n = 1). Average math course grade includes first-semester coursework in algebra, precalculus, calculus, and statistics.

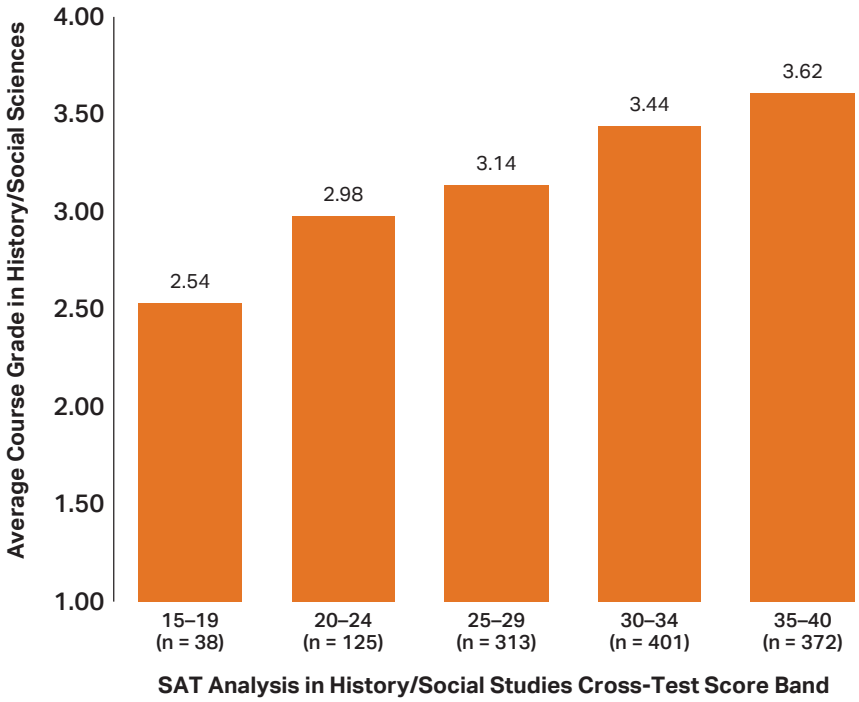
Figure 7.5: Relationship between SAT Analysis in Science cross-test scores and course grades in the same domain



Note: Results based on fewer than 15 students aren't reported (e.g., score band 10–14, n = 0). Average science course grade includes first-semester coursework in natural sciences, health sciences, and engineering.

Figure 7.6 shows the relationship between SAT Analysis in History/Social Studies cross-test scores and average first-semester, credit-bearing college course grades in history (e.g., world history, U.S. history, European history) and social sciences (e.g., anthropology, economics, government, geography, psychology) coursework. This graph depicts a clear positive relationship between SAT Analysis in History/Social Studies cross-test scores and grades in matching college courses. For example, those students with an SAT Analysis in History/Social Studies cross-test score of 20–24 have an average matching college course grade of 2.98, whereas those students with an SAT Analysis in History/Social Studies cross-test score of 35–40 have an average matching college course grade of 3.62.

Figure 7.6: Relationship between SAT Analysis in History/Social Studies cross-test scores and course grades in the same domain



Note: Results based on fewer than 15 students aren't reported (e.g., score band 10–14, n = 3). Average history/social studies course grade includes first-semester coursework in history and social sciences.

Discussion

This pilot predictive validity study of the new SAT allowed for a preliminary look at the relationship between new SAT scores and grades in the first year of college. Across a diverse sample of first-year students at 15 four-year institutions, the results of this pilot study showed that new SAT scores remain as predictive of college success as old SAT scores. This is important to note as the redesign of the SAT was first and foremost focused on more closely aligning the content and skills tested on the SAT with those content and skills that research indicates are critical for college success. In making these important changes to the test, that the strong predictive validity was also maintained is an important accomplishment of the redesign.

In addition, this study showed that new SAT scores improve the ability to predict college performance above HSGPA alone—and more so than in studies utilizing the old SAT. In other words, while the SAT and HSGPA are both measures of a student's previous academic performance that are strongly related to FYGPA in college, they also tend to measure somewhat different aspects of academic performance and therefore complement each other in their use in college admission and the overall prediction of FYGPA.

Finally, the examination of the relationships between the SAT section scores and cross-test scores with grades in the matching coursework domain(s) in college shows a strong positive relationship, suggesting that the redesigned SAT is sensitive to instruction in English/language arts, math, science, and history/social studies. Just as one would expect, higher SAT section or cross-test scores are associated with higher course grades in the matching academic field in college.

Given that this study was conducted with a pilot form of the test, a smaller sample, and students who may be less motivated to perform at their best, it will be important to replicate the study findings with a large, nationally representative sample now that the new SAT is operational. The College Board will be launching such a study, examining students in the entering college class of fall 2017, the first full cohort to be admitted to college with the new SAT. These students will complete one year of college and then in fall 2018 and the year that follows, we will be able to study the relationship between redesigned SAT scores and first-year college performance. We will continue to track students through college so that relationships between redesigned SAT scores and longer-term outcomes such as persistence, completion, and cumulative GPA, can also be studied.

7.5 Measuring and Monitoring College Readiness with the SAT

Background Information

Having demonstrated that new SAT scores remain as predictive of college success as old SAT scores; we now turn our attention to the notion of college readiness. As stated in Section 1.1, one of the primary intended uses of the SAT is to evaluate and monitor a student's college and career readiness. Each assessment in the SAT Suite of Assessments has an associated set of metrics called the college and career readiness benchmarks. These benchmarks are tools developed to help administrators, teachers, parents, and students understand whether students have mastered the knowledge, skills, and understandings needed to be successful in college. Given the role of the SAT as an assessment of students' college readiness, the importance of the benchmarks cannot be overstated. As explained in Section 1.2, the failure of a majority of students to meet the benchmark on the old SAT was one of the factors that precipitated a "call to action" to redesign the test. This section describes the processes that were used to develop and validate the SAT benchmarks for college and career readiness.

Benchmarks

Benchmarks are calculated separately for the SAT Math section and the SAT Evidence-Based Reading and Writing (ERW) section. In order for a student to be considered college and career ready, they must meet both the Math and ERW benchmarks. Previous research has

demonstrated that students who are considered college and career ready have higher overall college grades, higher retention rates,[4] and higher graduation rates compared to students who didn't meet the college and career readiness benchmark (Mattern, Shaw, & Marini, 2013; Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011).

Benchmark Criteria

The SAT College and Career Readiness Benchmarks represent the minimum scores on the Math and ERW sections that are associated with having a high likelihood of earning at least a C in relevant credit-bearing, introductory college-level courses. That is, the Math benchmark is set to be the minimum score a student can earn on the Math section of the SAT to have a 75% chance of earning at least a C in first-semester, credit-bearing, college-level courses in college algebra, statistics, precalculus, and/or calculus. If a student completed more than one related Math course in the first semester, the course with the lowest grade is used. Similarly, the ERW benchmark is set to be the minimum score a student can earn on the ERW section of the SAT to have a 75% chance of earning at least a C in first-semester, credit-bearing, college-level courses in history, literature, social science, and/or writing. If a student completed more than one related English course in the first semester, the course with the lowest grade is used. Using the lowest course grade for both Math and ERW ensures that students who meet or exceed the benchmark score are prepared to succeed in all of their first semester, introductory credit-bearing college courses.

Data

The college and career readiness benchmarks were empirically calculated and subsequently validated using two samples. The first sample, used to calculate the benchmarks, contains the SAT scores of students who graduated high school in 2009 that were matched to their postsecondary transcript records from two-year and four-year institutions. The ERW benchmark was calculated using approximately 116,000 first-semester course grades from 224 postsecondary institutions. The Math benchmark was calculated using approximately 53,000 first-semester course grades from 209 postsecondary institutions. Analyses to validate the calculated benchmarks were conducted using a second sample containing postsecondary enrollment information for over 1.1 million SAT takers.[5]

Analyses

Analyses were conducted using the old SAT that was administered between March 2005 and February 2016, prior to the redesign. Logistic regression analyses were used to calculate benchmark scores in ERW[6] and Math associated with a 75% probability of earning a C or higher in relevant first-semester, college-level coursework. In order to make Sample 1 representative of the population of two- and four-year college-going students, the data from the first sample were weighted so that the percentage of course enrollments at two- and four-year institutions approximated that of student enrollment in the population.[7] Additionally, the four-year institutions in our sample were weighted by the mean SAT score of the entering class so that our sample of four-year institutions resembled the population of four-year institutions in terms of selectivity.

---

[4] Retention is defined as students who return to their initial postsecondary institution in subsequent years.

[5] This data set used postsecondary enrollment records obtained from the National Student Clearinghouse (NSC).

[6] Critical reading and writing scores from the SAT administered prior to the redesign were combined to obtain the ERW benchmark score.

[7] Data on the population were obtained from the National Center for Education Statistics (NCES).

Results

Initial calculations, based on the SAT administered prior to the redesign yielded a Math benchmark score of 500 and an ERW score of 850. The Math score was based on the 200–800 scale and the ERW score, which combined critical reading and writing, was based on a 400–1600 scale. When placed on the more familiar 200–800 scale, the ERW benchmark score would be between 420 and 430. These scores were then concorded to the new SAT scale,[8] yielding concorded benchmark scores of 530 in Math and 480 in ERW. The benchmark scores from the SAT are shown in Table 7.6.

Table 7.6: The SAT College and Career Readiness Benchmarks

SAT Section	College Courses	SAT Benchmark
ERW	Literature, history, social science, and writing	480
Math	College algebra, statistics, precalculus, and calculus	530

Validity Evidence

Overall College and Career Readiness

To collect evidence to validate the inferences made from the college and career readiness benchmarks, analyses were conducted comparing the postsecondary outcomes of students who met the benchmark and the outcomes of students who didn't meet the benchmarks. Table A-7.45 in Appendix 7: Validity displays the results of analyses comparing retention, performance, and graduation for students enrolled in four-year institutions, in two-year institutions, and those students in two-year institutions who are also enrolled in workforce training programs. Results indicate that students who met both the ERW and Math benchmarks had higher retention rates, higher grades in college, and graduated at higher rates than those students who did not meet the benchmark.

Individual Benchmark Attainment

Tables A-7.46 and A-7.47 in Appendix 7: Validity indicate that meeting the separate benchmarks is related to performance in relevant college coursework. In both ERW and Math, students who met the benchmark were less likely to need remediation and had higher grades in related coursework than students who didn't meet the benchmark. These relationships existed for students in both two-year and four-year institutions.

Intended Uses

The college and career readiness benchmarks on the assessments in the SAT Suite provide a measure of the likelihood that students are ready to succeed in introductory, credit-bearing coursework during their first semester in college. The benchmarks are intended to be used by policymakers, administrators, educators, and parents to monitor the academic progress of a student or groups of students as they prepare for college and careers. When necessary, interventions can be introduced to help students get on track to graduating high school prepared for college-level work.

[8] The current, redesigned version of the SAT launched in March 2016.

7.6 Year-Over-Year Growth and PSAT-Related Benchmarks

Monitoring Progress Over Time

Section 7.5 describes how the SAT benchmarks were calculated. The SAT benchmarks provide evidence about the likelihood that a graduating high school senior will be successful in first-semester college courses. The benchmarks also provide information to those who are not college and career ready about academic areas that need strengthening. Such information is only useful, however, if it is received early enough to plan and execute interventions to get a student who has fallen behind back on track by graduation from high school. For this reason, it's helpful to have benchmarks that reach back into earlier grades.

Here we describe how the benchmarks for the earlier assessments in the SAT Suite were derived from the SAT benchmarks. Essentially, we computed average growth from one assessment to the next by finding the difference between average performances on consecutive assessments. Using the estimates of average yearly growth and back mapping from the SAT benchmarks allowed the determination of benchmarks from 11th grade back to 8th grade.

Creating PSAT Benchmarks

The PSAT/NMSQT, PSAT 10, and PSAT 8/9 launched in fall 2015, ahead of the SAT. In order to develop the benchmarks ahead of the assessment launch, we relied on the relationships among pre-2015 assessments to gauge estimates of growth from one grade to the next so that we could establish benchmarks for 8th through 11th grades.

Data

Two data sets were used for these analyses. First, we used data for 403,412 students from the 2014 matched cohort, which includes students who graduated from high school in 2014, who took the old PSAT/NMSQT in sophomore and junior years, and who took the SAT. Second, we used data from 3,704 eighth graders who completed the ReadiStep assessment,[9] as well as the old PSAT/NMSQT as part of a 2008 national pilot used to establish the original ReadiStep scale. We weighted both samples to make them representative of the national population of students at each grade level.[10] As a final step in preparing the data set for analyses, we concorded the scores at each grade level (8th through 11th) to the new scale of the SAT Suite.

Analyses

Average performance on ERW and Math was computed for each year in the data sets. We then estimated average yearly growth by subtracting the average performance in each year from the average performance in the subsequent year. We then computed each year's benchmarks by subtracting the average growth from the subsequent year's benchmark. For example, to compute the PSAT/NMSQT (11th grade) benchmarks, we started with the SAT benchmarks of 480 for ERW and 530 for Math. We subtracted the average growth between

[9] ReadiStep, designed for eighth graders, has been replaced by PSAT 8/9.

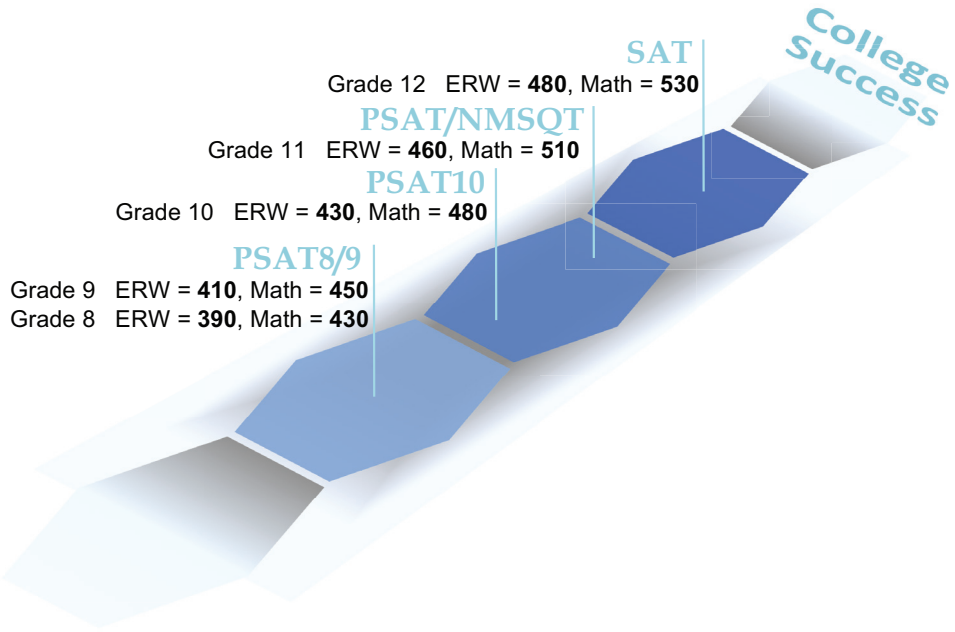
[10] Both of these data sets were weighted to the High School Longitudinal Study of 2009 (HSLS:09), a nationally representative sample, using region of the country, gender, ethnicity, and highest level of parental education (National Center for Education Statistics, n.d.).

11th and 12th grades (20 points for each section) from these benchmarks to obtain the 11th-grade benchmarks of 460 for ERW and 510 for Math. We repeated the process for the remaining grades (10th back to 8th).

Results

As expected, we found consistent growth on the estimated SAT Suite scores from year to year. The growth ranged from 20 to 30 points each on the ERW and Math section scores. Subtracting each year's growth from the previous year's benchmarks, starting with the SAT benchmarks (see Section 7.5), yielded the benchmarks shown in Figure 7.7.

Figure 7.7: Benchmarks for the redesigned SAT Suite of Assessments for Evidence-Based Reading and Writing and Math



Validity Evidence

To begin to collect validity evidence for the growth-based benchmarks, we looked for (1) at least as many people attaining the benchmark in a given year as achieved the benchmark in the previous year and (2) at least 75% attainment of the next-level benchmark for all of those attaining the benchmark at any given level. To examine the SAT attainment numbers, we used the 2015 cohort of 1,698,521 test takers. To examine the percentage attaining the benchmark at earlier levels, we used the 2015 Matched Cohort, which includes all students who graduated in 2015 and who took at least one assessment in the SAT Suite. The total data file had data for 3,024,776 test takers.

Table 7.7 shows the percentages of test takers attaining each section benchmark (ERW or Math) in each year. The table also shows the percentage of those attaining both benchmarks and thus deemed on track to be college and career ready.

Table 7.7: The Percentage of Test Takers Meeting Each Benchmark

Assessment	N	Section	Benchmark	% Meeting
SAT	1,698,521	ERW	480	71.6
		Math	530	53.5
		Both		50.0
PSAT/NMSQT	1,592,242	ERW	460	69.2
		Math	510	54.0
		Both		49.9
PSAT 10	1,580,161	ERW	430	65.3
		Math	480	48.9
		Both		44.7
PSAT 8/9 (9th)	14,797	ERW	410	58.4
		Math	450	37.8
		Both		34.0
PSAT8/9 (8th)	227,994	ERW	390	55.0
		Math	430	38.9
		Both		33.8

The table shows that an estimated 55% of eighth-grade PSAT 8/9 test takers meet the ERW benchmark and about 39% meet the Math benchmark. About 34% of test takers meet both. Among ninth-grade PSAT 8/9 test takers, 58% meet the ERW benchmark, 38% meet the Math benchmark, and 34% meet both. Thus, at least as great a percentage of test takers meet the ninth-grade benchmark as meet the eighth-grade benchmark. An even greater percentage meets each subsequent benchmark, a pattern that would indicate that those on track tend to stay on track.

Table 7.8 gives more direct evidence that at least 75% of those meeting the benchmark in a given year met the benchmark the following year. The table shows the number of students taking an assessment in a given year as well as in the subsequent year. The first column gives the initial grade level; the second column gives the section (ERW, Math, or both); the third column gives the total number of students taking the test in that grade as well as the subsequent test the following year; while the fourth and fifth columns show the percentage and the number meeting the benchmark that first year. The last column shows, of those meeting the benchmark the first year, the percentage of test takers who also met the benchmark the next year. We can see that of those meeting the ERW or the Math benchmarks or both in any given year, more than 80% met the benchmark the following year.

Table 7.8: The Percentage of Test Takers Who Met the Benchmark in a Given Year Who Also Met the Benchmark the Following Year

Grade	Section	Count	Percent Meeting Benchmark	Number Meeting Benchmark	Percent Meeting Next Benchmark
Nine	ERW	8,146	67.6	5,503	87.8
	Math	8,146	46.2	3,765	82.2
	Both	8,146	42.6	3,469	82.1
Ten	ERW	891,131	75.8	675,281	84.5
	Math	891,131	59.9	533,432	85.9
	Both	891,131	56.1	499,651	82.6
Eleven	ERW	678,066	68.3	463,373	97.0
	Math	678,066	63.0	427,135	88.5
	Both	678,066	54.9	372,005	91.3

In comparing Table 7.8 to 7.7, we notice two things. First, only about half of those who take an assessment in a given year take the subsequent assessment the following year. Second, the percentage of those in Table 7.8 who meet the benchmark is higher than the corresponding percentage in Table 7.7. Thus, the test takers who take the subsequent assessment tend to be stronger academically. Still, few people who are shown to be on track for college readiness on an assessment fall off track on the next assessment. This result gives us confidence that the benchmarks for the redesigned assessment will collectively give the score user consistent information about readiness.

REFERENCES

Achieve, Inc., The Education Trust, and Thomas B. Fordham Foundation. (2004). *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?* (pp. 163–189) New York, NY: Guilford Press.

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

Barry, C. L., & Niu, S. (2013). *Examining the impact of state-driven AP enrollment policies on student access to rigorous course work* (College Board Research Note 2013-4). New York, NY: College Board.

Beard, J., & Marini, J. P. (2015). *Validity of the SAT for predicting first-year grades: 2012 SAT validity sample* (College Board Statistical Report No. 2015-2). New York, NY: College Board.

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

Brennan, R. L. (2004). Manual for LEGS Version 2.0. (3). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes* (CASMA Monograph No. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. Retrieved from http://www.uiowa,edu/~casma

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

College Board. (2006). *College Board standards for college success: English language arts*. New York, NY: Author. Retrieved from http://www.asainstitute.org/conference2008/featuredsessions/collegeboard-english-language-arts_cbscs.pdf

College Board. (2009). *ACT and SAT concordance tables*. (RN-40). New York, NY: Author. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2009-40-act-sat-concordance-tables.pdf

College Board. (2011). *Guidelines on the uses of College Board test scores and related data*. New York, NY: Author. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/guidelines-on-uses-of-college-board-test-scores-and-data.pdf

College Board. (2014). The redesigned SAT: Evidentiary foundation. In *Test specifications for the redesigned SAT* (pp. 24–38). New York, NY: Author.

College Board. (2015a). *Compare SAT specifications*. New York, NY: Author. Retrieved from https://collegereadiness.collegeboard.org/sat/inside-the-test/compare-current-new-specifications

College Board. (2015b). ScorEquate [Computer software]. New York, NY: Author.

College Board. (2015c). *Scores and scales & student-produced response (SPR) business rules*. New York, NY: Author. Retrieved from https://wiki.collegeboardnewmedia.org/display/SR/Scores+and+Scales

College Board. (2015d). *Services for students with disabilities. Ensuring accommodations on College Board exams*. New York, NY: Author. Retrieved from https://www.collegeboard.org/students-with-disabilities

College Board. (2016a). *College Board program results*. New York, NY: Author. Retrieved from https://www.collegeboard.org/program-results

College Board. (2016b). *Scaling for the SAT Suite of Assessments*. New York, NY: Author.

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.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando FL: Harcourt Brace.

Dorans, N. J. (2000). *Distinctions among classes of linkages* (RN-11). New York, NY: College Board. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-2000-11-distinctions-classes-linkages.pdf

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124.

Gal, I. (2002). Adults' statistical literacy: Meaning components. Responsibilities. *International Statistical Review*, *70*(1), 1–25.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, *12*, 971–988.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: Praeger.

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

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

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

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

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

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

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. New York, NY: Springer Publishing.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurements for score scales. *Journal of Educational Measurement*, *29*, 285–307.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate student's success. *Science*, *315*(5815), 1080–1081.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh*, *A-62*, 28–30.

Mattern, K. D., Kobrin, J. L., Patterson, B. F., Shaw, E. J., & Camara, W. J. (2009). Validity is in the eye of the beholder: Conveying SAT research findings to the general public. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 213–240). Charlotte, NC: Information Age.

Mattern, K. D., & Patterson, B. F. (2014). *Synthesis of recent SAT validity findings: Trend data over time and cohorts* (College Board Research in Review 2014-1). New York, NY: College Board. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2014/6/Synthesis-of-Recent-SAT-Validity-Findings.pdf

Mattern, K. D., Shaw, E. J., & Marini, J. (2013). *Does college readiness translate to college completion?* (College Board Research Note RN 2013-9). New York, NY: College Board.

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.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*(2), 128–165.

Micciche, L. R. (2004). Making a case for rhetorical grammar. *College Composition and Communication*, *55*(4), 716–737.

Miles, J. N. V., & Shevlin, M. E. (2001). *Applying regression and correlation: A guide for students and researchers*. London, United Kingdom: Sage.

Moses, T., & Golub-Smith, M. (2011). *A scaling method that produces scale score distributions with specific skewness and kurtosis* (Research Memorandum No. RM-11-04). Princeton, NJ: Educational Testing Service.

Moses, T., & Kim, Y. K. (2017). Stabilizing conditional standard errors of measurement in scale score transformations. *Journal of Educational Measurement*, *54*, 184–199.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York, NY: 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 for Education Statistics. (n.d.). *High school longitudinal study of 2009*. Retrieved from https://nces.ed.gov/surveys/hsls09/

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. (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.

Pearson, K. (1902). On the mathematical theory of errors of judgment, with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *198*, 235–299.

Roderick, M., Coca, V., & Nagaoka, J. (2011). High school effects in shaping urban students' participation in college application, choice, and enrollment. *Sociology of Education*, *84*(3), 178–211.

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

Schumacker, R., & Muchinsky, P. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions*, *10*, 479.

Seburn, M., Frain, S., & Conley, D. T. (2013). *Job training programs curriculum study*. Eugene, OR: Educational Policy Improvement Center.

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

Shaw, E. J., Marini, J. P., Beard, J., Shmueli, D., Young, L., & Ng, H. (2016). *The redesigned SAT pilot predictive validity study: A first look* (College Board Research Report No. 2016-1). New York, NY: College Board.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*. *50*, 99–104.

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

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

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating*. New York, NY: Springer Publishing.

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.

Wyatt, J., Kobrin, J., Wiley, A., Camara, W. J., & Proestler, N. (2011). *SAT benchmarks: Development of a college readiness benchmark and its relationship to secondary and postsecondary school performance* (College Board Research Report No. 2011-5). New York, NY: College Board. Retrieved from http://research.collegeboard.org/rr2011-5

Wyatt, J., Wiley, A., Camara, W. J., & Proestler, N. (2011) . *The development of an index of academic rigor for college readiness* (College Board Research Report No. 2011-11). New York, NY: College Board.

GLOSSARY

Accommodations/Test Accommodations: Adjustments that don't alter the assessed construct which are applied to test presentation, environment, content, format (including response format), or administration conditions for particular test takers and that are embedded within assessments or applied after the assessment is designed. Tests or assessments with such accommodations, and their scores are said to be *accommodated*. Accommodated scores should be sufficiently comparable to unaccommodated scores so that they can be aggregated together.

Accuracy: The closeness of a measured or calculated value to a standard, known, or true value. Compare to *precision*.

Administration: The presentation of a test to one or more test takers. Test administration involves item format, mode of presentation, timing or pacing of the test, and the resulting record of the administered test. When the test is given under the same conditions and using the same instructions for all test takers, the test and administration are said to be standardized.

Alpha Reliability (Coefficient): An internal consistency reliability coefficient based on the number of parts into which a test is partitioned (e.g., items, subtests, or raters), the interrelationships of the parts, and the total test score variance. The coefficient rests on the assumption that all parts measure the same construct or dimension. Also called Cronbach's alpha and, for dichotomous items, *KR20*. *See* Internal Consistency Coefficient.

Alpha Reliability (Cronbach's): *See* Alpha Reliability (Coefficient).

Alpha Reliability (Stratified): A modification of alpha reliability appropriate for tests made up of several parts or strata, each measuring a somewhat different dimension or construct; when the alpha reliability coefficient can be computed for each part.

Alternate Forms: Two or more versions of a test that are considered interchangeable in that they measure the same constructs in the same ways, are built to the same content and statistical specifications, and are administered under the same conditions using the same directions.

Assessment: Any systematic method of obtaining information, used to draw inferences about characteristics of people, objects, or programs; a systematic process to measure or evaluate the characteristics or performance of individuals, programs, or other entities for purposes of drawing inferences; sometimes used synonymously with *test*.

Automated Test Assembly: A technique for creating test forms that involves the use of mathematical optimization algorithms to select items from an item pool to build test forms adhering to a set of preprogrammed content and statistical specifications and other constraints.

Benchmark: A standard or point of reference against which things may be compared or assessed. In educational measurement, a level of test performance indicating a given level of student achievement (e.g., proficient) relative to some set of specific learning goals.

College and Career Readiness: The level of preparation a student needs in order to be ready to succeed in first-year, credit-bearing courses in postsecondary education (two-year, four-year) or in workforce training programs.

Concordance: In linking test scores for tests that measure similar constructs, the process of relating a score on one test to a score on another, so that the scores have the same relative meaning for a group of test takers.

Conditional Standard Error of Measurement: The standard deviation of measurement errors that affect the scores of test takers at a specified test score level.

Construct: The concept or characteristic that a test is designed to measure.

Construct-Irrelevant Variance: Variance in test taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation.

Content-Oriented Validity: Evidence based on test content that supports the intended interpretation of test scores for a given purpose. Such evidence may address issues such as the fidelity of test content to performance in the domain in question and the degree to which test content representatively samples a domain, such as a course curriculum or job.

Correlation: The extent that two variables are related. If high scores on one variable are related to high scores on the second variable, then the relationship is positive. The correlation coefficient ranges from −1.00 (perfect negative relationship) to +1.00 (perfect positive relationship). A zero correlation coefficient indicates no statistical relationship between the two variables. Most correlations of test scores and measures of academic success are positive. The higher the correlation is, the better the prediction.

Criterion: In an educational testing setting, a value associated with a person (e.g., a performance measure such as future grade point average) that we wish to estimate or predict by use of a test score. In psychometric analyses of item discrimination, the criterion is often the test score.

Differential Item Functioning (DIF): For a particular item in a test, a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses or, in some cases, different rates of choosing various item options.

Differential Validity: Evidence indicating differences in the relationship between a test and criterion by subgroup (e.g., gender, race/ethnicity, best language).

Difficulty (item): A statistic indicating the proportion of test takers in a given population who will answer an item correctly. One index of item difficulty is the *p*-value.

Discrimination (item): A statistic indicating the change in the proportion of test takers answering an item correctly associated with a given increase in the knowledge level of those test takers. An index of the ability of a test item to differentiate among test takers of differing achievement levels.

Domain: The set of interrelated attributes (e.g., behaviors, attitudes, values) that are included under a construct's label.

English Language Learner (ELL): An individual who isn't yet proficient in English, generally whose first language isn't English. Related terms include *English learner* (EL), *limited English proficient* (LEP), *English as a second language* (ESL), and *culturally and linguistically diverse*.

Equated Forms: Alternate forms of a test whose scores have been related through a statistical process known as equating, which allows scale scores on equated forms to be used interchangeably.

Equating: A process for relating scores on alternate forms of a test so that they have essentially the same meaning. The equated scores are typically reported on a common score scale.

Fairness: Fairness refers to the validity of test score interpretations for their intended use(s) across students in all pertinent subgroups. It is a multifaceted and overarching concept in measurement covering the equitable treatment of all test takers in a test administration as well as equal measurement quality across subgroups and populations.

Field Test: A test administration used to check the adequacy of testing procedures and the statistical characteristics of new test items or new test forms. *See also* Pilot Study.

Focal Group (DIF): In evaluating test items, the group of test takers of interest, which is compared to a reference group, with the purpose of determining whether membership in the focal group might be related to frequency of answering an item correctly, even after controlling for knowledge and skills. *See* Differential Item Functioning, Reference Group (DIF).

Growth: Student academic progress over time. Often measured by comparing test scores on the same students over time.

Intercorrelation: *See* Correlation.

Internal Consistency: *See* Internal Consistency Coefficient.

Internal Consistency Coefficient: An index of the reliability of test scores derived from the statistical interrelationships among item responses or scores on separate parts of a test. *See* Alpha Reliability (Coefficient).

Internal Structure: In test analysis, the factorial structure of item responses or subscales of a test.

Item: A statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task. *See* Prompt.

Item Bank: A repository for storing and maintaining test questions (items) and the information pertaining to them.

Item Characteristics: Characteristics and measures of the properties (content or statistical) of a test item. *See also* Differential Item Functioning (DIF), Difficulty (Item), Discrimination (Item).

Item Pool: A collection of test questions (items) available for a specific usage purpose, such as the construction of test forms.

Norms: Statistics or tabular data that summarize the distribution or frequency of test scores for one or more specified groups, such as test takers of various ages or grades, usually designed to represent some larger population, referred to as the *reference population*.

Operational (Administration): The presentation of a test to a group of test takers, with the goal of using the test scores to inform an interpretation, decision, or action related to the test's intended purpose. *See* Administration, Operational Use.

Operational (Data): Information collected from an *operational administration*.

Operational (Forms): Forms of a test designated for use in an *operational administration*.

Operational (Items): Test items designed for use in an *operational administration*.

Operational Use: The actual administration of a test form to intended test takers, after initial test development has been completed, to inform an interpretation, decision, or action, based in part or wholly on test scores.

Passage: A text source upon which test items are based.

Percentile Rank (Related to Norms): The rank of a given score based on the percentage of scores in a specified score distribution that are below the score being ranked.

Pilot Study: A study that is conducted to inform operational assessment plans, principles, uses, or procedures.

Pilot Test: A study that informs assessment plans, principles, or procedures to achieve broad public goals. *See also* Field Test.

Precision: The closeness of several measured values to each other. The consistency with which some value is measured. Compare to *accuracy*.

Predictive Validity: Evidence indicating how accurately test data collected at one time can predict criterion scores that are obtained at a later time.

Pretest: An administration of items or test forms to obtain statistics and/or proof of concept before the items are used in an *operational administration*.

Prompt: The question, task, stimulus, or instruction that elicits a test taker's response.

Raw Score: A score on a test that is calculated by counting the number of correct answers, or more generally, a sum or other combination of item scores.

Reference Group (DIF): In evaluating test items, the group of test takers to which other groups are compared, with the purpose of determining whether group membership is related to frequency of answering an item correctly, even after controlling for knowledge and skills. *See* Differential Item Functioning, Focal Group (DIF).

Reference Group (Norms): The population of test takers to which individual test takers are compared through the test norms. The reference population may be defined in terms of test taker age, grade, clinical status at the time of testing, or other characteristics. *See* Norms.

Reliability: The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group. *See* Precision.

Restriction of Range: Reduction in the observed score variance of a test taker sample, compared with the variance of the entire test taker population, as a consequence of constraints on the process of sampling test takers.

Sample: A selection of a specified number of entities, called sampling units (test takers, items, etc.), from a larger specified set of possible entities, called the *population*.

Sampling Error: When estimating a population value (e.g., the mean) from a sample from that population, the difference between the sample estimate and the actual population value.

SAT Suite of Assessments: Introduced as part of the College Board Readiness and Success System, a system designed to make it easier for students to navigate a path through high school, college, and career. The SAT Suite comprises the SAT, PSAT/ NMSQT, PSAT 10, and PSAT 8/9, all of which focus comprehensively on the few durable skills that evidence shows matter the most for college and career success.

Scale: The system of numbers, and their units, by which a value is reported on some dimension of measurement. In testing, the set of items or subtests used to measure a specific characteristic (e.g., a test of verbal ability or a scale of extroversion-introversion).

Scale Score: A score obtained by transforming raw scores. Scale scores are typically used to facilitate interpretation.

Scaling: The process of creating a scale or a scale score to enhance test score interpretation by placing scores from different tests or test forms on a common scale or by producing scale scores designed to support score interpretations. *See* Scale.

Score Tier: For the SAT Suite of Assessments, one of the several levels at which scale scores are reported, including total score; section scores (Evidence-Based Reading and Writing [ERW], Math [MSS]); test scores (Reading [R], Writing and Language [WL], Math [MTS]); cross-test scores (Analysis in Science [SCI], Analysis in History/Social Studies [HSS]); and subscores (Words in Context [WIC], Command of Evidence [COE], Expression of Ideas [EOI], Standard English Conventions [SEC], Heart of Algebra [HOA], Problem Solving and Data Analysis [PSD], Passport to Advanced Mathematics [PAM]).

Security: *See* Test Security.

Speededness: The extent to which test takers' scores depend on the rate at which work is performed as well as on the correctness of the responses. The term isn't used to describe tests of speed.

Standard Error of Measurement: The standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions. Because such data generally cannot be collected, the standard error of measurement is usually estimated from group data.

Standard Error of Sampling: *See* Sampling Error.

Standardized Administration: In test administration, maintaining a consistent testing environment and conducting tests according to detailed rules and specifications, so that testing conditions are as similar as possible for all test takers on the same and multiple occasions.

Statistical Specifications: *See* Test Specifications.

Subgroup: A subset of students belonging to the same group that may be of interest for test reporting, analyses, or research. For example, many test analyses include information on students by gender and racial/ethnic subgroups.

Test: An evaluative device or procedure in which a systematic sample of a test taker's knowledge, skills, and/or abilities in a specified domain is obtained and scored using a standardized process. Sometimes synonymous with *assessment*; sometimes references a *score tier*.

Test Blueprint: *See* Test Specifications.

Test Design: The process of developing detailed specifications for what a test is to measure and the content, cognitive level, format, and types of test items to be used.

Test Development: The process through which a test is planned, constructed, evaluated, and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for the test's intended purpose.

Test Form: A set of items or exercises that meet the requirements of the specifications for a testing program. Many testing programs use test forms that are intended to be parallel, meaning that each form is built according to the same specifications but with some or all of the test items unique to each form. *See* Alternate Forms.

Test Security: Administration procedures and analyses implemented to protect the content of a test from unauthorized release or use and to preserve the integrity of the test scores so they are valid for their intended use.

Test Specifications: Documentation of the purpose and intended uses of a test, as well as of the test's content characteristics, format, length, time limits, psychometric characteristics (of the items and the test overall), delivery mode, administration, scoring, and score reporting.

Text Complexity: The level of challenge a text provides to a reader. Measurement of text complexity typically falls in one of three types: qualitative measures (e.g., levels of meaning, structure, clarity, knowledge demands); quantitative measures (e.g., word frequency and length, sentence length); and reader/text considerations (e.g., student knowledge and motivation).

Universal Design: An approach to assessment development that attempts to maximize the accessibility of a test for all of its intended test takers.

User (Group): A group of test takers who may not represent a well-defined reference population. Instead, the term may refer to the self-selected group who took a particular test (perhaps during a specified time frame).

User (Norms): Descriptive statistics (including percentile ranks) for a group of test takers that doesn't represent a well-defined reference population, for example, all persons tested during a certain period of time, or a set of self-selected test takers. *See* Norms.

Validity: The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed.

Vertical Scaling: In test linking, the process of relating scores on tests that measure the same construct but differ in difficulty. Typically used with achievement and ability tests with content or difficulty that span a variety of grade or age levels.