CollegeBoard SAT®

Scaling for the SAT Suite of Assessments

Scaling for the SAT Suite of Assessments

YoungKoung Kim is a senior psychometrician in the College Board's Psychometrics department and is the co-editor of the monograph.

Tim Moses is the Robert L. Brennan Chair of Psychometric Research in the College Board's Psychometrics department and is the editor of the monograph.

Michael J. Kolen is a Psychometric Advisor at the College Board.

Amy Hendrickson is a senior director in the College Board's Psychometrics department.

About the College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of over 6,000 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success—including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators, and schools. For further information, visit: **collegeboard.org**.

Preface

In December 2014, the College Board conducted a large study to develop the scales of the redesigned SAT Suite of Assessments. The purpose of this monograph is to describe the methodology and scale development process in more detail so that the monograph can serve as a supplement to the documents that contain more general descriptions of the redesigned SAT Suite of Assessments (College Board, 2014, 2017). Chapters are specifically devoted to the general overview of the SAT Suite of Assessments scaling (Chapter 1); study design, data cleaning, and weighting (Chapter 2); the SAT scales (Chapter 3); the vertical scales for the PSAT-related tests (Chapter 4); and the subscore scales (Chapter 5).

This scaling monograph is the final product of the redesigned SAT Suite of Assessments Scaling project. The Scaling project could not be completed without help from several people. We would like to acknowledge the encouragement and inspiration we received from College Board Leadership, especially Kevin Sweeney, Jack Buckley (former College Board Senior Vice President), and Cyndie Schmeiser. We also deeply appreciate the valuable advice and guidance from the College Board Psychometric Advisors, Michael Kolen and Robert Brennan, throughout the Scaling project.

This scaling monograph was produced not only by the chapter authors and editors, but also with the help and constructive comments of many reviewers from both inside and outside the College Board. We are grateful to them for helping us to improve the writing and presentations contained within this work. From the College Board, Rosemary Reshetar, Judit Antal, Jay Happel, Paula Cunningham, Jane Dapkus, Andrew Courchane, and Kelcey Edwards provided tremendous insight. Robert Brennan and Won-Chan Lee of the University of Iowa offered invaluable input, as did the National Merit Scholarship Corporation. Finally, we would like to thank Mark Syp for the editorial reviews and Gail Mitnik for her project management assistance.

Tim Moses
YoungKoung Kim
May, 2017

Contents

Tables

Figures

CHAPTER 1

Overview of Scaling—Rationale and Goals

Michael J. Kolen

The primary purpose of Chapter 1 is to provide an overview of the rationale and goals for the development of the score scales for the SAT Suite of Assessments, which includes the SAT®, PSAT/NMSQT® and PSAT™ 10, and PSAT™ 8/9 assessments. The chapter begins with a general discussion of the purposes of the assessments. This discussion includes a description of test content with an emphasis on content alignment across assessments. The SAT Suite of Assessments is also compared to previous SAT-related assessments. The chapter continues with an overview of the goals and processes used to develop score scales for the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9.[1]

The intent of Chapter 1 is to provide an overview. Subsequent chapters describe the process of developing score scales in greater detail.

Purpose of the Assessment Redesign

In this section, the purpose of the redesign is considered, and the SAT Suite of Assessments is briefly compared to the previous SAT-related assessments. Test content alignment across assessments in the SAT Suite is also briefly described.

SAT Suite of Assessments Versus Previous SAT-Related Assessments. Based on input from members, partner organizations, and postsecondary and K–12 experts, the College Board identified three assessment challenges that the SAT Suite of Assessments is intended to address. First, the SAT Suite is intended to provide a more comprehensive and informative picture of student readiness for college-level work and workforce training while sustaining the ability of the test to predict college success. Second, the SAT Suite is intended to focus on the knowledge, skills, and understandings that research indicates are essential for college and career readiness and success. Third, the SAT Suite is intended to reflect, through its questions and tasks, the kinds of meaningful, engaging, and challenging work that students undertake in the best high school courses, so as to strengthen the bond between assessment and instruction.

In order to achieve these goals, the Reading Test and Writing and Language Test focus on words in context and command of evidence. The Math Test has a deep focus on fewer topics that are essential for college readiness. The questions on the Reading Test, Writing and Language Test, and Math Test are grounded in the real world and are directly related to work performed in college and careers. Across all components of the assessment, students are asked to apply their reading, writing, language, and math skills to answer questions in

[1] 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.

science, history, and social studies contexts. In addition, there is no penalty for guessing, which encourages students to give the best answer for every question. Compared to earlier SAT-related assessments, the SAT Suite has more scores that are intended to support a wider range of purposes (College Board, 2014, 2017).

Test Content Alignment Across Assessments. Another major goal was for test content to be aligned across the assessments. Such content alignment is intended to allow for direct assessment of student growth across the assessments. In order for the content to be aligned, tests in the SAT Suite measure the same skills and knowledge in ways that are appropriate for students at different grade levels. As students progress through high school, the tests are intended to keep pace, matching the scope and difficulty of work being done in the classroom (For more details, refer to College Board, 2017).

On the Reading Test, as students advance from the PSAT 8/9 to the PSAT/NMSQT and PSAT 10 to the SAT, they encounter longer passages, more questions, and more questions pertaining to informational graphics. On the Writing and Language Test, students are asked to make increasingly sophisticated choices in vocabulary, sentence structure, organization, tone, and factual support. On the Math Test, students will see more multistep math problems and more problems that require the use of complicated concepts and equations. The number of math problems on the Math Test also increases from the PSAT 8/9 to the SAT: the number of student-produced response questions increases and the proportion of multiple-choice questions decreases.

Scale Scores and Derived Scores

The SAT reports a variety of scale scores to examinees. *Scale scores* on the SAT are found by transforming the number of items correctly answered on a given set of items (*raw scores*) to scale scores by applying the appropriate raw-to-scale score conversions. Table 1.1 lists the 12 SAT scale scores along with the score range, score increment, and intended mean scale score for the SAT cohort. This SAT cohort is described later in this chapter and in detail in Chapter 2.

Derived scores are found as a sum or weighted sum of scale scores. Table 1.2 lists the three SAT derived scores along with the score range, score increment, and mean score for the SAT cohort. The Math Test score is found by dividing the Math section score by 20, which leads to scale scores that range from 10 to 40 with a score increment of .5. The Evidence-Based Reading and Writing section score is found by summing the Reading Test score and the Writing and Language Test score, and multiplying the sum by 10. The total score is found by summing the Math section score and the Evidence-Based Reading and Writing section score.

SAT scores (Total, Section, Test, Cross-Test, and subscores) are intended to be used by all K–12 educators to assess and improve college and career readiness and success for high school students, as well as by higher education institutions for admission and placement purposes.

The PSAT/NMSQT and PSAT 10 and PSAT 8/9 contain the same scores as the SAT, except that the subscore Passport to Advanced Math is included only on the PSAT/NMSQT and PSAT 10 and the SAT. Also, as described later, the score ranges on the PSAT/NMSQT and PSAT 10 and PSAT 8/9 differ from those on the SAT.

Table 1.1: SAT Scale Scores

	Score Range	Score Increment	Intended SAT Cohort Mean
Section Score			
Math (MSS)	200–800	10	500
Test Score			
Reading (R)	10–40	1	25
Writing and Language (WL)	10–40	1	25
Cross-Test Score			
Analysis in History/Social Studies (HSS)	10–40	1	25
Analysis in Science (SCI)	10–40	1	25
Subscore			
Command of Evidence (COE)	1–15	1	8
Words in Context (WIC)	1–15	1	8
Expression of Ideas (EOI)	1–15	1	8
Standard English Conventions (SEC)	1–15	1	8
Heart of Algebra (HOA)	1–15	1	8
Passport to Advanced Mathematics (PAM)	1–15	1	8
Problem Solving and Data Analysis (PSD)	1–15	1	8

Table 1.2: SAT Derived Scores

	Score Range	Score Increment	SAT Cohort Mean
Test Score			
Math (MTS)	10–40	0.5	25
Section Score			
Evidence-Based Reading and Writing (ERW)	200–800	10	500
Total Score (Total)	400–1600	10	1000

SAT Scaling

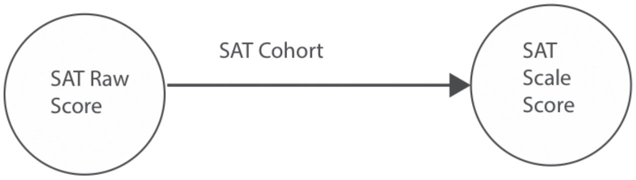
The SAT scores were developed using data from a *2014 scaling study*, in which groups of 11th- and 12th-grade students that were intended to be nationally representative were administered the SAT. As discussed in more detail in Chapter 2, the data from this study were weighted using statistical weighting procedures to be representative of graduating seniors who took the SAT over the last four years. This weighted sample is referred to as the *SAT cohort*.

For the 12 scale scores, the score range and score increment were set as indicated in Table 1.1. The goals for constructing the raw-to-rounded scale score conversion tables for the initial test form that was used in the 2014 scaling study were as follows:

1. A number-correct raw score of “none correct” was always converted to the lowest scale score, and a number-correct raw score of “all correct” was always converted to the highest scale score.
2. Scale score increments that were used are shown in Table 1.1.
3. The mean scale scores for the SAT cohort were set equal to the values shown in Table 1.1.
4. The standard deviation of the Math section score for the SAT cohort was set to be approximately 100. The standard deviations of the test scores and cross-test scores shown in Table 1.1 were set to be approximately 5. The standard deviations of the subscores shown in Table 1.1 were set to be approximately equal to one another.
5. The conditional standard errors of measurement for each scale score shown in Table 1.1 were set to be approximately equal along the score scale. In this way, the standard error of measurement for a scale score will be approximately equal for all examinees, which is intended to facilitate test score interpretation.
6. The raw-to-scale score conversions were set to minimize gaps and many-to-one conversions in the rounded scale score conversion tables.
7. For the purpose of maintaining the score scales over test forms and over time using equating procedures, raw-to-unrounded scale score conversions were also developed. The rounded raw-to-scale score conversion can be obtained when the unrounded scale scores are rounded with respect to the scale score increment.
8. The standard deviation of the Math section score and the derived Evidence-Based Reading and Writing section score were set to be approximately equal.

Figure 1.1 illustrates the scaling process. The circles in this figure are used to represent the scores involved in the scaling. The line with the arrow indicates that the SAT raw scores are being linked (transformed) to SAT scale scores. The text above the line with the arrow indicates that the scaling, based on the eight goals above, is conducted using the SAT cohort. The specific psychometric methods used to develop these score scales are described in Chapter 3. The derived scores shown in Table 1.2 are computed from the scale scores shown in Table 1.1.

Figure 1.1: SAT Scaling



Scaling and Vertical Scaling for the PSAT/NMSQT and PSAT 10 and PSAT 8/9

Scaling of PSAT/NMSQT and PSAT 10 and PSAT 8/9 was also based on data collected in the 2014 scaling study. The scalings obtained from this study were based on a group of 10th graders who took the PSAT/NMSQT and PSAT 10, and a group of 9th graders who took the PSAT 8/9. The data were statistically weighted to form *nationally representative groups*. Specific procedures are described in Chapters 2 and 4.

Subscores on the PSAT/NMSQT and PSAT 10 and the PSAT 8/9 were constructed independently of those for the SAT, using data from the nationally representative groups to have a mean of 8 and a scale score range of 1 to 15. For the PSAT/NMSQT and PSAT 10, data from a group of 10th graders was used to construct the scale. For the PSAT 8/9, data from the grade 9 nationally representative group was used to construct the scale. As with the SAT subscores, the standard deviations were set to be approximately equal and the conditional standard errors of measurement were intended to be approximately equal along the score scale.

The other scale scores for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 were constructed using vertical scaling procedures that make use of the *scaling test design* (Kolen & Brennan, 2014). In the 2014 scaling study, each examinee took a complete test along with a randomly assigned *scaling test* from one of the following five subjects: Math, Reading, Writing and Language, Analysis in History/Social Studies, or Analysis in Science. These scaling tests were developed to represent the content and statistical characteristics of a combined SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 test. That is, the scaling tests were developed to cover the domain of content over the SAT Suite of Assessments.

The same scaling test was administered to examinees in grades 9, 10, and 11 in the 2014 scaling study. Thus, a common scaling test was administered to students in each of the grades. The use of such a scaling test allows for the comparison of students from different grades on a common set of items. By examining data on the scaling test for nationally representative (NR) groups of examinees at each grade, a direct estimate of how much growth is exhibited from one grade to the next can be made. For example, the difference in scaling test means for representative samples of grade 10 and grade 11 examinees provides direct evidence of how much, on average, students grow from grade 10 to grade 11 on the entire domain of content. Because growth is being assessed over the domain of content for the SAT Suite of Assessments, Kolen and Brennan (2014) refer to the process as being consistent with a *domain definition of growth*.

Vertical scaling was conducted using the scaling test and chained equipercentile linking (Kolen & Brennan, 2014). In this method, number-correct scores on the PSAT/NMSQT and PSAT 10 (or PSAT 8/9) are linked to scores on the scaling test using equipercentile procedures based on the NR sample of 10th graders who took the PSAT/NMSQT and PSAT 10 in the 2014 scaling study. The scores on the scaling test are linked to the SAT scale scores using chained equipercentile linking based on the NR group of 11th graders who took the SAT in the 2014 scaling study. The two linking functions are chained together to provide a raw-to-scale score conversion table for the PSAT/NMSQT and PSAT 10.

By applying the chained equipercentile method, scores on the PSAT/NMSQT and PSAT 10 are converted to scores on the SAT scale that was already developed. Note that the linking makes use of the NR samples, and not the SAT cohort that was used to develop the SAT scale.

The vertical scaling process is illustrated in Figure 1.2. The circles in this figure represent test scores, and the text within the figure indicates the test score represented by the associated circle. Lines with arrows indicate a linking of test scores and the text beside the arrow indicates the group of examinees used to conduct the linking. By following the circles and arrows, it can be seen that scores on the PSAT 8/9 are linked to scaling test raw scores using the NR group of 9th-grade examinees. The scaling test raw scores are linked to SAT scale scores using the NR group of 11th graders. These two linking functions are chained together to link PSAT 8/9 raw scores to SAT scale scores. A similar description applies to linking PSAT/NMSQT and PSAT 10 raw scores to SAT scale scores.

Because the PSAT 8/9 test and PSAT/NMSQT and PSAT 10 tests assess lower level content and are intended to be easier than the SAT, the minimum and maximum scores are lower on the PSAT 8/9 and the PSAT/NMSQT and PSAT 10 than on the SAT for the test scores that were vertically scaled. The score ranges for the tests that were vertically scaled are shown in Table 1.3.

PSAT/NMSQT and PSAT 10 and PSAT 8/9 use derived scores that are calculated in the same way as the derived scores used with the SAT. The score ranges and score increments for the derived scores are shown in Table 1.4. Because the scores that are used to calculate the derived scores are vertically scaled and the same process is used to calculate the derived scores for the SAT, the PSAT/NMSQT and PSAT 10, and the PSAT 8/9 the derived scores are considered to be *vertically aligned*.

Figure 1.2: Vertical Scaling

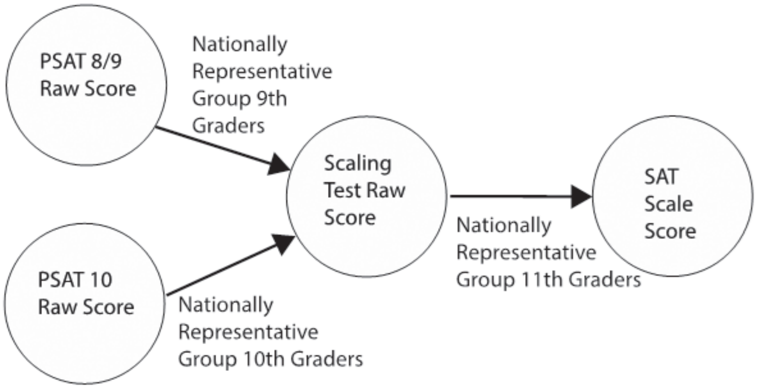

PSAT 8/9 Raw Score
Nationally Representative Group 9th Graders
Scaling Test Raw Score
SAT Scale Score
Nationally Representative Group 11th Graders
PSAT 10 Raw Score
Nationally Representative Group 10th Graders

Table 1.3: PSAT 8/9 and PSAT/NMSQT and PSAT 10 Vertical Scale Scores

	PSAT 8/9 Score Range	PSAT/NMSQT and PSAT 10 Score Range	Score Increment
Section Score			
Math (MSS)	120–720	160–760	10
Test Score			
Reading (R)	6–36	8–38	1
Writing and Language (WL)	6–36	8–38	1
Cross-Test Score	6–36	8–38	
Analysis in History/Social Studies (HSS)	6–36	8–38	1
Analysis in Science (SCI)	6–36	8–38	1

Summary of SAT and PSAT Scaling

The score scales for the SAT were constructed to serve two general purposes. The section scores and total score are intended to be used by colleges for admission and admission-related purposes and by middle school and high school teachers to evaluate students' overall progress toward college readiness. The section scores use a score scale that ranges from 200 to 800. The test scores, cross-test scores, and subscores are intended for use by high schools to assess and improve students' college and career readiness and success.

The scores for the PSAT/NMSQT and PSAT 10 and the PSAT 8/9 have direct parallels to the SAT scores (except that one of the Math subscores is not included on the PSAT 8/9). The inclusion of vertically scaled test scores and vertically aligned derived scores on the PSAT/NMSQT and PSAT 10 and the PSAT 8/9 are intended to facilitate the assessment of student growth.

Table 1.4: PSAT-Related Derived Scores

	PSAT 8/9 Score Range	PSAT/NMSQT and PSAT 10 Score Range	Score Increment
Test Score			
Math (MTS)	6–36	8–38	0.5
Section Score			
Evidence-Based Reading and Writing (ERW)	120–720	160–760	10
Total Score (Total)	240–1440	320–1520	10

The use of an approximately constant standard error of measurement on all of the SAT scores is intended to facilitate score interpretation by examinees, in that a single standard error of measurement can be used when interpreting the amount of measurement error in a score.

This chapter provided an overview of the process for score scales. The following chapters provide much more detail. Chapter 2 describes the 2014 Scaling Study. Chapter 3 describes the SAT scaling and Chapter 4 the PSAT scaling. Chapter 5 focuses on the scaling of subscores.

BIBLIOGRAPHY/REFERENCES

College Board. (2014). *Test specifications for the redesigned SAT*. New York, NY: The College Board.

College Board. (2017). *SAT Suite of Assessments technical manual: Characteristics of the SAT*. New York, NY: The College Board.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking. Methods and practices.* (3rd ed.). New York, NY: Springer.

CHAPTER 2

Scaling Study Design—Sampling, Test Administration, Data Cleaning, and Weighting

Amy Hendrickson and Tim Moses

Procedures for obtaining the data used to establish the scales for the redesigned SAT Suite of Assessments involved defining target populations for recruitment of the study participants, recruitment and test administration activities, and post-processing of the test data by cleaning for motivation approximations and weighting for demographic approximations. These procedures were developed and implemented to be consistent with the goals of the scaling for the SAT Suite, namely to develop SAT scales on examinee groups that reflect recent cohorts of graduating seniors, and to develop PSAT/NMSQT and PSAT 10 and PSAT 8/9 scales on examinee groups that reflect grade-specific nationally representative groups (see Chapter 1). This chapter describes the steps for defining recruitment targets, recruitment activities, data cleaning, and weighting for the scaling study data.

Target Population Definitions for the Recruitment of the Scaling Study Samples

The targets for recruitment for the scaling samples were nationally representative samples of high school 9th, 10th, 11th, and 12th graders. National representation was operationalized as a sample of schools that matched the Common Core of Data 2011–2012 (Keaton, 2012) information regarding U.S. high schools in terms of College Board region (New England, Middle States, Southern, Midwestern, Southwestern, Western), urbanicity (City, Suburb, Town, Rural), public or private, and percentage of enrolled students receiving free or reduced-price lunch.[1] The College Board partnered with Westat to provide a target list of schools, and backup schools, that if recruited would provide nationally representative samples of 15,000 students in 9th, 10th, 11th, and 12th grade.

The grade levels of these samples corresponded to the scaling targets for the PSAT/NMSQT and PSAT 10 (10th graders) and the PSAT 8/9 (9th graders), which were determined using analyses described in Chapter 4. The PSAT/NMSQT and PSAT 10 and PSAT 8/9 were linked to the SAT using the samples of students who took those tests and a group of 11th graders who took the SAT (NRSAT, for the nationally representative SAT sample used in the vertical scaling).

[1] Due to conflicting interest with other College Board research studies, schools from a few states were excluded from the target sampling plan.

Sample Design and Data Collection Procedures

Recruitment. Using the sample of schools provided by Westat, ETS data collection services staff recruited schools to participate in the study. Schools may have been included in the sampling plan for only one grade level, but they were able to choose to have one, two, or all grade-level students participate. Schools participating in the SAT and/or PSAT/NMSQT and PSAT 10 administration had to be registered at an official College Board test center for the study. All schools also had to participate in a training webinar prior to the study.

A total of 457 schools participated, with approximately 138 public schools and 12 private schools per grade level. Schools were rewarded $15 for each student that came to the administration, and students who tried to complete the test were rewarded a $50 gift card and a free SAT practice test.

Test Forms and Survey. Three forms of the SAT, three forms of the PSAT/NMSQT and PSAT 10, and one form of the PSAT 8/9 assessments were created for the study. These forms were designed to meet the content specifications and statistical specifications for these assessments (College Board, 2014, 2017). Additionally, five scaling tests were created, one each in Reading, Writing and Language, Math (composed of two separately timed sections), Analysis in History/Social Studies, and Analysis in Science. These scaling tests were developed to represent the content and statistical characteristics of a shortened version of a combined SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 test. That is, the scaling tests were developed to cover the domain of content over the SAT Suite of Assessments. Students also completed a survey in addition to the test forms (Appendix A).

Test Administration. The test administration window occurred between December 9, 2014, and February 20, 2015. Students completed a full-length assessment corresponding to their grade level. Thus, 11th and 12th graders completed an SAT form; 10th and 11th graders completed a PSAT/NMSQT and PSAT 10 form (not the same 11th graders that completed the SAT form); and 9th graders completed a PSAT 8/9 form. In addition, each test taker completed one of the separately timed scaling tests, which was an hour in length. Separate Math calculator and Math no calculator scaling tests were developed but were administered together to the same students, as separately timed tests, and administered in classrooms separate from the other scaling tests.

The full-length forms of each assessment were spiraled within appropriate grade levels for the administration. Scaling tests were spiraled in conjunction with the full version test to achieve randomly equivalent groups completing the scaling tests. All students of a given grade level completed the full-length tests together. Then, those students who had been randomly assigned the Math scaling tests were moved to other rooms to allow for separate timing of the two sections and for students to use their calculators when allowed. The booklets containing the Math scaling tests were printed with an identifying label and color so that the proctors knew to move those examinees to another room. Total required student time was 4 to 5 hours. Because of the differences in timing for the SAT, PSAT/NMSQT and PSAT 10, and the PSAT 8/9, the 9th, 10th, and 11th graders were tested in separate rooms.

Target Populations for the Scaling

Target Population for SAT Scaling. One of the goals for establishing the SAT scales was to use data reflective of recent SAT cohorts (see Chapter 1). This goal suggests that the target population for the SAT scales was an average of recent SAT cohorts (i.e., annual groups of college-bound graduating seniors). This group can be most clearly described in terms of its demographic construction. In particular, Table 2.1 shows the proportions of demographic subgroups for SAT cohorts from 2011–2014. The sample for this SAT group consisted primarily of 11th- and 12th-grade students, and was weighted to reflect the proportions of demographic subgroups shown in Table 2.1. The weighting procedure is described in more detail at the end of this chapter.

The sample dataset used for the scaling of the SAT tests was cleaned with the intent to approximate the motivation level of examinee performance expected by the SAT cohort of college-bound seniors. Data cleaning levels were considered based on percentages of questions completed on the Reading, Math, and Writing tests; on answering at least one student-produced response item on the operational Math Test (not the scaling test); and also on self-reported motivation on the survey administered with the SAT test (Appendix A). The implications of these levels and the final choice used to approximate plausible motivation levels of the historical SAT cohort are described in Chapter 3.

Target Population for the Vertical Scaling. The target population for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 vertical scales were nationally representative high school students. Nationally representative targets were defined as proportions of demographic subgroups defined in recent National Center for Education Statistics (NCES) reports (Table 2.2; Bitterman, Gray, & Goldring, 2013; Broughman & Swaim, 2013; Keaton, 2012; Keaton, 2013). Grade levels for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 examinees were also part of the target population definition, where nationally representative 10th and 9th graders were the target grades for PSAT/NMSQT and PSAT 10, and PSAT 8/9, respectively (though 11th graders taking the PSAT/NMSQT and PSAT 10 were also considered; Chapter 4).

The target population for the SAT students used to establish the vertical scales for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 was nationally representative. This SAT sample is designated as NRSAT. NRSAT included only 11th graders (though 11th and 12th graders were considered; Chapter 4). Except for the grade levels, NRSAT had a target population that was the same one used for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 samples, but a different target population from the one used to set the SAT scales. The NRSAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 samples were weighted to reflect the demographic proportions of nationally representative subgroups described in Table 2.2.

The target for analysis was 3,000 examinees per vertical scaling group (i.e., 15 vertical scaling groups based on three tests and two cross-tests for the SAT, for the PSAT/NMSQT and PSAT 10, and for the PSAT 8/9) for a maximum of 45,000 examinees. A preliminary model for the vertical scaling study, using Reading as an example, is shown in Table 2.3. (Five of these tables exist, one for each vertical scale.) The model for the full study is illustrated in Table 2.4.

Table 2.1: Target Population Based on Recent SAT Cohorts for SAT Scaling

Subgroup	2011–2014 SAT Cohort %	Subgroup	2011–2014 SAT Cohort %
Females	**53**	**Desired College Degree**	
Juniors	**33**	Specialized training or certificate program	1
Ethnicity		Two-year associate of arts or sciences degree	1
American Indian or Alaska Native	1	Bachelor's degree	26
Asian, Asian American, or Pacific Islander	12	Master's degree	26
Black or African American	13	Doctoral or related degree	19
Mexican or Mexican American	7	Other	1
Puerto Rican	2	Undecided	13
Other Hispanic, Latino, or Latin American	9	**Interested in Attending Type of College**	
White	51	Four-year college or university	79
Other	4	Two-year community or junior college	11
Region		Vocational/technical school	2
MRO	6	Undecided	5
MSRO	25	**Mother's Highest Ed**	
NERO	8	Grade school	3
SRO	21	Some high school	5
SWRO	11	High school diploma or equivalent	16
WRO	22	Business or trade school	2
First Language		Some college	14
English Only	69	Associate or two-year degree	9
English and another language	16	Bachelor's or four-year degree	24
Another language	12	Some graduate or professional school	3
Best Language		Graduate or professional degree	13
English Only	75		
English and another language	19		
Another language	4		

Table 2.1 continued on next page

Table 2.1 continued from previous page

Subgroup	2011–2014 SAT Cohort %	Subgroup	2011–2014 SAT Cohort %
Father’s Highest Ed		A–	18
Grade school	3	B+	17
Some high school	6	B	15
High school diploma or equivalent	18	B–	8
Business or trade school	3	C+	5
Some college	12	C	3
Associate or two-year degree	5	C–	1
Bachelor’s or four-year degree	21	D+	0
Some graduate or professional school	2	D	0
Graduate or professional degree	16	E or F	0
Average High School GPA		**High School Math: AP/Honors**	
A+	6	AP/Honors	30
A	18	**High School English**	
		High School English: AP/Honors	34

The sample datasets used for the vertical scaling of the PSAT/NMSQT and PSAT 10 and the PSAT 8/9 tests, and the NRSAT dataset also used in these scalings were cleaned to approximate the expected motivation levels of nationally representative 11th- (NRSAT), 10th- (PSAT/NMSQT and PSAT 10), and 9th- (PSAT 8/9) grade examinees. Data cleaning levels were considered based on percentages of questions completed on the Reading, Math, and Writing Tests; on answering at least one student-produced response item on the operational Math Test (not the scaling test); on absolute differences between standardized test or cross-test scores versus the scaling test scores less than 3; and also on self-reported motivation on the survey administered with the PSAT/NMSQT and PSAT 10 and PSAT 8/9 tests (Appendix A). The implications of these levels and the final choice used to approximate nationally representative motivation levels are described in Chapter 4.

Table 2.2: Target Population Based on Recent NCES Targets Used for Vertical Scaling

Subgroup	Nationally Representative
Post Secondary Intension	72
Female (vs. Male)	50
Private/Public	5/95
	Private/Public
American Indian	<1/1
Asian	<1/4
Black	<1/13
Hispanic	1/19
White	3/55
Other or Missing	<1/2
MRO	1/21
MSRO	1/14
NERO	<1/4
SRO	1/21
SWRO	<1/11
WRO	1/23
Rural	<1/26
Suburban	2/31
Town	<1/11
Urban	2/27

Table 2.3: Intended Implementation of the Scaling Test Sampling Used for the Vertical Scaling (for the Reading Tests)

Group	Main Test	Scaling Test	Grade Levels	Targetted N
1	Full NRSAT (3 spiraled forms)	Reading Scaling	11	3,000
2	Full PSAT/NMSQT and PSAT 10 (3 spiraled forms)	Reading Scaling	10	3,000
3	Full PSAT 8/9 (1 form)	Reading Scaling	9	3,000

Table 2.4: Intended Implementation of the Scaling Test Sampling Used for the Vertical Scaling (full study)

Group	Main Test	Scaling Test	Grade Levels	Targetted N
1	Full NRSAT (3 spiraled forms)	Reading, Writing/Language, Math (cal and no calc), Science, Social Studies	11	15,000
2	Full PSAT/NMSQT and PSAT 10 (3 spiraled forms)	Reading, Writing/Language, Math (cal and no calc), Science, Social Studies	10	15,000
3	Full PSAT 8/9 (1 form)	Reading, Writing/Language, Math (cal and no calc), Science, Social Studies	9	15,000

Weighting

The data collected for the scaling were considerably different from the target populations. Specifically, the target populations for the school-level recruitment differed from the student-level demographic compositions of the target populations for setting the scales. To adjust for these differences, the datasets obtained after cleaning for motivation were weighted to more closely resemble the target populations of interest for the scalings. Because the target populations were defined in terms of proportions of student-level demographic subgroups (Tables 2.1 and 2.2), these subgroup proportions present targets that might be approximated through weighting.

The weighting of the scaling study data involved obtaining student-level weights referred to as post stratification weights or "case weights" for each student, such that the entire sample more closely represents the distribution of background variables for a targeted population (Cochran, 1977; Valliant, Dever, & Kreuter, 2013). The weighting model of interest is a loglinear model with the following form,

$$\ln(w_k) = \Sigma_{r=0}^{R} D_{rk}\beta_r = \mathbf{D_k}\boldsymbol{\beta}, \qquad (2.1)$$

where student k's weight, w_k, is obtained as a function of R measured variables (plus a D_{0k} variable set to 1), $\mathbf{D_k} = [D_{0k}\ D_{1k} \ldots \ldots .\ D_{Rk}]$ (Haberman, 1984, 2014). The $D_{0k} = 1$ and β_0 produce an intercept that ensures the w_k's sum to 1. The complete set of model parameters, $[\beta_0\ \beta_1\ \beta_2 \ldots \beta_R]^t = \boldsymbol{\beta}$, can be estimated such that the weighted means of the D_{rk}s approximate pre-specified target values, $\Sigma_k D_{rk}\hat{w}_k \approx Target_r$, $r = 1$ to R. Weighting applications can address desires to approximate proportions of demographic subgroups in the sample data. For example, the D_{rk}'s may be a set of dummy variables indicating whether students are male ($D_{males,k}$), Asian ($D_{Asians,k}$), from the Northeast region of the United States ($D_{Northeast,k}$), etc., and the w_k's are obtained such that a weighted sample simultaneously reflects desired proportions of males ($\Sigma_k D_{males,k}\hat{w}_k \approx Target_{males}$), Asians ($\Sigma_k D_{Asians,k}\hat{w}_k \approx Target_{Asians}$), and students from the Northeast ($\Sigma_k D_{Northeast,k}\hat{w}_k \approx Target_{Northeast}$). The use of the loglinear model allowed for weighting to reflect the target proportions of a large list of several overlapping background variables.

Although the weighting applications used for the scaling study datasets generally improved the representativeness to the scaling study target populations (Chapters 3 and 4), several issues were encountered that affected the results. The success of matching the target proportions of several demographic variables was affected by the following:

- **Sample Sizes:** The successfulness of matching target proportions was influenced by sample size, in particular overall sample size and sample size for the subcategories of the background variables. Convergence problems can occur when large loglinear models are fit to datasets with small sample sizes. When sample data have sample sizes that are either very small or zero for particular subcategories of variables, weighting for these categories may not be possible with the data currently used and choices must be made to either collapse those categories or distribute them across other subcategories.
- **Model Fitting Criteria and Algorithms:** The weighting models being fit were based on matching the target proportions of several background variables. Algorithms for estimating the weighting models included Newton algorithms (Haberman, 1984) and Generalized Reduced Gradient algorithms. The algorithm that was used produced converged solutions that matched the target proportions as closely as possible while also meeting overall scaling goals.
- **Responses to Background Variables:** Responses to the background variables were not perfectly accurate, in that they tended to reflect limited or missing information for some categories either due to the data gathering process or in survey responses.
- **Extreme Weights:** When sample sizes for subcategories were small and target proportions were not, examinee case weights could be extreme. Concerns for extreme weights are that they may inflate the variability of estimates (Valliant, Dever, & Kreuter, 2013) such as the scaling results obtained from the weighted data. Ad hoc trimming procedures have been described and were considered with respect to reducing extreme weights and approximating target proportions.
- **Limited Information on Target Distributions for Variables of Interest:** Weighting results would have been improved if accurate target distributions were available on several important variables. Some of the more important background variables were examinee motivation for college-bound and nationally representative students, and college intention for nationally representative students.

The weighting algorithms were applied in consideration of all of the above issues, and, most importantly, to produce results that would support the goals of the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 scalings (see Chapter 1). In several situations, multiple options were available for approximating the target populations for scaling, such as combining or reassigning categories with sample sizes that were small or zero, producing extreme weights to approximate target proportions for several background variables, or to address background variable responses that might be nonexistent or inaccurate. To address these issues, multiple weighting solutions were considered and used to produce scaling solutions, and these solutions were compared and evaluated with respect to scaling goals such as alignment of the means of the section scores, test scores, and cross-test scores, etc. Choices were made for weighting solutions that were most successful in terms of fulfilling multiple scaling goals, while approximating target populations for scaling as closely as possible.

BIBLIOGRAPHY/REFERENCES

Bitterman, A., Gray, L., & Goldring, R. (2013). *Characteristics of public and private elementary and secondary schools in the United States: results from the 2011–12 schools and staffing survey* (NCES 2013–312). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [April 15, 2015] from http://nces.ed.gov/pubsearch.

Broughman, S. P., & Swaim, N. L. (2013). *Characteristics of private schools in the United States: results from the 2011–12 private school universe survey* (NCES 2013–316). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [April 15, 2015] from http://nces.ed.gov/pubsearch.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley.

College Board. (2014). Test Specifications for the Redesigned SAT. New York, NY: The College Board.

College Board. (2017). *SAT Suite of Assessments technical manual: Characteristics of the SAT.* New York, NY: The College Board.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics, 12,* 971–988.

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (ETS Research Memorandum RM-14-01). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*(3), 254–273.

Keaton, P. (2012). *Public elementary and secondary school student enrollment and staff counts from the common core of data: school year 2010–11* (NCES 2012-327). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [April 15, 2015] from http://nces.ed.gov/pubsearch.

Keaton, P. (2013). *Selected statistics from the common core of data: school year 2011–12* (NCES 2013–441). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [April 15, 2015] from http://nces.ed.gov/pubsearch.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). Practical tools for designing and weighting survey samples. New York, NY: Springer.

CHAPTER 3

SAT Scaling—Characteristics of New SAT Scaling

YoungKoung Kim and Tim Moses

As discussed in Chapter 1, the purpose of scaling is to establish numerical systems that convey test performance. Better scales are the ones that support intended interpretations of test performance, which for the SAT involves the scale score systems summarized in Chapter 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. 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 chapter describes the goals for establishing the new SAT scales, the data collection for the scaling study, the scaling process, and the results (Standards 5.1–5.2, AERA/APA/NCME, 2014). Additional work for evaluating the scales is also discussed. This chapter focuses on the procedures of developing the section and test/cross-test scale scores, while the scaling process of the seven subscores is discussed in Chapter 5.

Goals for the Scales

The scale scores were established as conversions of the number-correct scores for 12 scores of the redesigned test. This process was based on goals consistent with how the scores were intended to be established. As discussed in Chapter 1, the scores were intended to be used by K–12 educators to assess and improve college and career readiness and success for high school students, as well as by higher education institutions for admission and placement purposes.

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 (*CSEM*s)[1] that are approximately constant and similar along the entire score range.

[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 due to sampling that reflects samples of examinees.

- All correct, maximum possible raw scores that convert to a highest obtainable scale score of 800.
- None correct, minimum possible raw scores that 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, and Writing and Language Test scores and Analysis in Science and Analysis in History/Social Studies cross-test 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.
- Conditional standard errors of measurement (*CSEM*s) that are approximately constant and similar along the entire score range.
- All correct, maximum possible raw scores that convert to a highest obtainable scale score of 40.
- None correct, minimum possible raw scores that convert to a lowest obtainable scale score of 10.
- Minimized gaps and many-to-one conversions in the raw-to-scale score conversion tables.

The scaling goals for the section, test, and cross-test scores are intended to support appropriate interpretations of SAT test performance. The scales are relatable across the section, test, and cross-test scores, 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 score interpretations and differentiations among test takers. Approximately equal measurement precision in terms of stabilized *CSEM*s 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 *CSEM*s. In addition, the SAT scales are intended to have similar standard deviations and similar distributional shapes across all scale scores of any given type.

Method

Data. Three SAT forms were administered in the 2014 Scaling Study (Tables 2.3 and 2.4). After evaluation of their statistical properties, one form was identified as the base form for the SAT and the data from the SAT base form was used for SAT scaling. In addition, a self-reported survey was administered with the SAT base form. There were 6,024 nationally recruited 11th- and 12th-grade examinees who took the SAT base form in the scaling study. Nationally recruited examinees were obtained from a list of high schools selected to achieve a nationally representative sample of high schools in terms of variables such as grades, states, and school type (private/public). From that group of examinees in the sampled high schools, we attempted to compose a sample that represents a typical SAT cohort group by identifying motivated examinees similar to the old SAT cohorts in terms of several background variables.

The scaling study sample was evaluated based on examinees' responses to the test items and the survey. Motivated and unmotivated examinees 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.[2] Other survey questions included in the SAT test administrations were used to identify a college-bound group, which was composed of 11th and 12th graders.[3] By considering completion rate, a survey question on examinees' motivation, grade level, and educational plans beyond high school, 10 samples based on different combinations of motivation screenings were initially examined.

Among the samples that were reviewed based on different degrees of screening, a sample of 4,346 examinees was selected because it was most desirable in terms of sample size, statistics, cohort representativeness, and scaling feasibility. The unweighted sample consists of 4,346 examinees who (1) were 11th or 12th graders; (2) completed 75% or higher percentage of the items on three tests—the Reading, the Math, and the Writing and Language Tests; (3) responded to at least one student-produced response (SPR) item on the Math Test; and (4) selected any of the response options, "I tried my best," "I gave moderate effort," or "I began by trying my best but I found the test very difficult" to the survey question about their effort.

Using the selected unweighted sample after screening for motivation, we attempted to create the weighted sample of 11th- and 12th-grade examinees. To compose the weighted sample, the weighting method described in Haberman (1984, 2014) 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, genders, examinees' college plans, College Board region, mother's and father's education, first and best language, GPA, and honors/AP coursework in Math and English.

The average subgroup percentages of these background variables for the old SAT cohort sample, which were used as the target percentages for weighting, are presented in Table 3.1. As shown in Table 3.1, 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%, while 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 seemed to be successfully achieved.

[2] Survey question 1 asked "Rate the level of effort that you gave while completing this test." The response options were "1 = I tried my best", "2 = I gave moderate effort", "3 = I gave little effort", "4 = I began by trying my best, but then I found the test very difficult; by the end of the test, I was no longer putting in much effort."

[3] Survey Question 8 asked, "What is the highest level of education you plan to complete beyond high school?" The response options were "I do not plan to pursue further education after high school"; "Specialized training or certificate program"; "Two-year associate of arts or associate of sciences degree"; "Bachelor's degree"; "Master's degree"; "Doctoral or related degree"; "Other"; and "Undecided."

Table 3.1: Unweighted Scaling, Weighted Scaling, and SAT Cohort Samples by Subgroups

Subgroup	Unweighted Scaling Sample %	Weighted Scaling Sample %	2011–2014 SAT Cohort %
Females	52	53	53
Juniors	67	33	33
Ethnicity			
American Indian or Alaska Native	1	1	1
Asian, Asian American, or Pacific Islander	10	12	12
Black or African American	14	13	13
Mexican or Mexican American	10	7	7
Puerto Rican	1	2	2
Other Hispanic, Latino, or Latin American	10	8	9
White	48	51	51
Other	3	4	4
Region			
MRO	14	8	6
MSRO	3	26	25
NERO	0	9	8
SRO	25	22	21
SWRO	36	12	11
WRO	23	23	22
First Language			
English Only	71	69	69
English and another language	16	16	16
Another language	13	12	12
Best Language			
English Only	71	75	75
English and another language	22	18	19
Another language	3	4	4

Table 3.1 continued on next page

Table 3.1 continued from previous page

Subgroup	Unweighted Scaling Sample %	Weighted Scaling Sample %	2011–2014 SAT Cohort %
Desired College Degree			
Specialized training or certificate program	3	1	1
Two-year associate of arts or sciences degree	4	1	1
Bachelor's degree	24	26	26
Master's degree	28	25	26
Doctoral or related degree	21	18	19
Other	2	1	1
Undecided	17	13	13
Interested in Attending Type of College			
Four-year college or university	80	79	79
Two-year community or junior college	19	10	11
Vocational/technical school	4	2	2
Undecided	8	5	5
Mother's Highest Ed			
Grade school	4	3	3
Some high school	5	5	5
High school diploma or equivalent	15	16	16
Business or trade school	2	2	2
Some college	12	14	14
Associate or two-year degree	7	9	9
Bachelor's or four-year degree	18	24	24
Some graduate or professional school	3	3	3
Graduate or professional degree	8	13	13
Father's Highest Ed			
Grade school	3	3	3
Some high school	6	6	6

Table 3.1 continued on next page

Table 3.1 continued from previous page

Subgroup	Unweighted Scaling Sample %	Weighted Scaling Sample %	2011–2014 SAT Cohort %
High school diploma or equivalent	15	18	18
Business or trade school	3	3	3
Some college	10	12	12
Associate or two-year degree	5	5	5
Bachelor’s or four-year degree	15	20	21
Some graduate or professional school	2	2	2
Graduate or professional degree	10	16	16
Average High School GPA			
A+	4	6	6
A	18	18	18
A−	19	18	18
B+	19	17	17
B	18	15	15
B−	9	8	8
C+	7	5	5
C	3	3	3
C−	1	1	1
D+	0	0	0
D	0	0	0
E or F	0	0	0
High School Math: AP/Honors			
AP/Honors	33	30	30
High School English			
High School English: AP/Honors	39	34	34

Scaling Methods. The two scaling methods considered were the arcsine transformation and a cubic transformation obtained from numerically stabilizing *CSEM*s estimated from the compound binomial model. For the arcsine transformation method, the raw scores, Y_j, $j = 0, \ldots, N_i$ are transformed using the following equation:

$$g(Y_j) = 0.5\left\{\sin^{-1}\left[\left(\frac{Y_j}{Ni+1}\right)^{\frac{1}{2}}\right] + \sin^{-1}\left[\left(\frac{Y_j+1}{Ni+1}\right)^{\frac{1}{2}}\right]\Bigg]\right\} \tag{3.1}$$

where N_i is the number of items on the test and $\sin^{-1}$ is the arcsine function (Kolen & Brennan, 2014). To obtain the desired mean and average conditional standard error of measurement (*CSEM*), the scale scores, $SC_{Y\,j,arcsine}$, can be found by linearly transforming the arcsine transformed scores as follows

$$sc_{Yj,arcsine} = \mu_{sc} + \frac{sem_{sc}}{\widehat{sem_g}}\left[g\left(Y_j\right) - \overline{g\left(Y_j\right)}\right] \tag{3.2}$$

where $\widehat{sem_g}$ is the estimated average standard error measurement (*SEM*) of the arcsine transformed scores, sem_{sc} is the desired average *SEM* of the scale scores, μ_{sc} is the desired scale score mean and $\overline{g(Y_j)}$ is the estimated mean of the arcsine transformed scores. For the SAT scaling, the compound binomial model was used to estimate $\widehat{sem_g}$, the *SEM* of the arcsine transformed scores:

$$\widehat{sem_g} = \sqrt{\frac{N_i - 2k}{4N_i^2 + 2N_i}} \tag{3.3}$$

where k is the Lord's k term (Lord, 1965) which is presented in a part of Equation 3.12. The details of the arcsine transformation stabilizing scale score *CSEM*s can be found in Kolen and Brennan (2014).

As an alternative method to the arcsine transformation method, a numerical approach to stabilizing *CSEM*s 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:

$$sc_{Yj,Cubic} = \delta_0 + \delta_1 Y_j + \delta_2 Y_j^2 + \delta_3 Y_j^3, \tag{3.4}$$

where the δ's are the polynomial coefficients (Moses & Golub-Smith, 2011). As originally described by Moses and Golub-Smith (2011), the values of the δ's are numerically solved to produce scale scores with prespecified skewness (γ_3) and kurtosis (γ_4). This was accomplished through minimizing the following function,

$$\left(\gamma_{sc(Y)3} - \gamma_3\right)^2 + \left(\gamma_{sc(Y)4} - \gamma_4\right)^2, \tag{3.5}$$

where $\gamma_{sc(Y)3}$ and $\gamma_{sc(Y)4}$ are the skewness and kurtosis of scale scores. 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 stability function for the scale score *CSEM*s defined as

$$\sum_{j=1}^{N_i} \left| CSEMsc_{Yj,Cubic} - CSEMsc_{Yj-1,Cubic} \right|. \tag{3.6}$$

$CSEM_{SC_{Y_j},Cubic}$ are obtained using the delta method as follows:

$$CSEMsc_{Y_j,Cubic} = \frac{\partial sc_{Y_j,Cubic}}{\partial Y} CSEM_{Y_j} = \left(\delta_1 + 2\,\delta_2 Y_j + 3\delta_3 Y_j^2\right) CSEM_{Y_j} \tag{3.7}$$

Compared to the arcsine transformation method, the cubic transformation is more explicitly obtained from the raw score *CSEM*s and also for *CSEM*s that are not necessarily based on binomial assumptions about dichotomous items. Scale score *CSEM*s can be estimated directly from the raw score *CSEM*s (Equation 3.7), which is simpler than computing conditional standard deviations of scale scores from measurement models (Kolen, Hanson, and Brennan, 1992). Finally, using cubic transformations, both *CSEM* stabilization and symmetry in scale score distributions can be simultaneously achieved if both are considered as important scaling goals. For the technical details of the cubic transformation, refer to Moses and Golub-Smith (2011) and Moses and Kim (2017).

Scaling Procedure. Because the ERW section score distribution reflected the bivariate distribution of the Reading and Writing and Language Test scores, its characteristics were more complicated to control than those of the Math section score. Thus, to manage the complex aspects of the ERW score distribution, scale score distributions for the Reading Test, the Writing and Language Test, and the derived ERW section scores were first established. Then, a scale score distribution for the Math section score that reflects the characteristics of the ERW section scores as similarly as possible was established. Scaling for the cross-test scores and subscores followed a similar process. The steps of SAT scaling procedures were as follows:

1. Examine Reading Test rounded scale score conversions corresponding to the various *CSEM* levels. Select an optimal *CSEM* level for the Reading Test scores by considering standard deviation, score gaps, and many-to-one conversions.[4]
2. Select the optimal *CSEM* level for the Writing and Language Test scores by considering standard deviation, score gaps, and many-to-one conversions. The standard deviation that was similar to the standard deviation of the Reading Test scores selected from Step 1 was preferred.
3. Examine the distribution of the ERW section scores. A standard deviation around 100 was preferred. In addition, review the estimated *CSEM* level for the ERW section scale scores.

[4] Given the raw correlation between Reading Test and Writing and Language Test scores (>0.83), the CSEM level that provided a standard deviation of slightly greater than 5.0 was preferred.

4. Examine the Math section rounded scale score conversions with the various *CSEM* levels. Select the *CSEM* that provides a standard deviation similar to the standard deviation of the ERW score distribution from Step 3. Examine score gaps and many-to-one conversions. If there is no optimal *CSEM* level that makes the scale score distribution similar between the two section scores and/or provides small numbers of score gaps and many-to-one conversions, go back to Step 1 or examine different levels of the *CSEM*s until the optimal *CSEM* level is found.
5. Examine the Analysis in Science cross-test scores for various *CSEM*s. Select the *CSEM* that provides the standard deviation that is similar to the standard deviation of the Reading and Writing and Language Test scores after Step 4. Also, consider the *CSEM* level that also provides a small number of score gaps and many-to-one conversions.
6. Select the *CSEM* level for the Analysis in History/Social Studies cross-test scores which produces a standard deviation similar to the standard deviations of the Analysis in Science cross-test scores from Step 5.

The scaling methods ultimately selected were those that produced the most similar means and standard deviations of the Math and ERW section scores, and the three test and two cross-test scores in the weighted data.

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. The following equations were used for linear interpolation outside the range:

$$
\begin{aligned}
Interpolated\_Upper_{Y} &= sc_{Y_{UpperBound}} + \left(sc_{y_{j'}} - sc_{Y_{UpperBound}}\right)\frac{\left(Y - Y_{UpperBound}\right)}{\left(Y_{j'} - Y_{UpperBound}\right)}, \\
Interpolated\_Lower_{Y} &= sc_{Y_{LowerBound}} + \left(sc_{y_{j'}} - sc_{Y_{LowerBound}}\right)\frac{\left(Y - Y_{LowerBound}\right)}{\left(Y_{j'} - Y_{LowerBound}\right)}.
\end{aligned}
\qquad (3.8)
$$

At the upper end, the scale score at the Y^{th} raw score, *Interpolated_Upper*$_Y$, was interpolated using the predetermined upper bound scale scores ($SC_{Y_{UpperBound}}$), the raw score associated with $SC_{Y_{UpperBound}}$ ($Y_{UpperBound}$), the scale scores that starts the interpolation ($SC_{Y_{j'}}$), and the raw score associated with $SC_{Y_{j'}}$ ($Y_{j'}$). Likewise, at the lower end, the scale scores at the Y^{th} raw score, *Interpolated_Lower*$_Y$ was interpolated using the predetermined lower bound scale scores ($SC_{Y_{LowerBound}}$), the raw score associated with $SC_{Y_{LowerBound}}$ ($Y_{LowerBound}$), the scale scores that starts the interpolation ($SC_{Y_{j'}}$) and the raw score associated with $SC_{Y_{j'}}$ ($Y_{j'}$).

Results

After conducting the iterative process for SAT scaling described earlier in this chapter, the raw-to-rounded scale score conversions for the SAT scale scores were developed. The arcsine transformation method was used to set the scales for the Reading Test score and Analysis in Science cross-test score, while the cubic transformation was used for the Math section score, the Writing and Language Test score, and Analysis in History/Social Studies cross-test score.

The Evidence-Based Reading and Writing (ERW) section scores were mathematically derived from the rounded Reading Test (R) and Writing and Language Test (WL) scale scores with the following mathematical expressions:

$$ERW = R \cdot 10 \ + WL \cdot 10 \qquad (3.9)$$

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

$$MTS = MSS/20 \qquad (3.10)$$

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

$$\text{Total} = \ ERW + MSS \qquad (3.11)$$

Table 3.2 shows summary statistics for the 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, test and cross-test, and subscores were almost identical to the target means of 500 for section scores and 25 for test and cross-test scores. In addition, all scale scores appeared to have similar average *CSEM*s and similar standard deviations across all scale scores of any given type. The *CSEM*s for the scales that were directly established from SAT scaling—Math section score, Reading Test score, Writing and Language Test score and two cross-test scores—were computed based on the method described in Kolen, Hanson, and Brennan (1992).

Strong true score models have been developed for the estimation of *CSEM*s given true scores (Kolen et al., 1992; Lord, 1965, 1969). The compound binomial distribution of the observed raw scores given the proportion of correct true scores τ can be approximated as,

$$\hat{\Pr}\left(Y = j \middle| \tau\right) = p_{Ni}(j) + \frac{N_i}{2} V_2 C_2(j), \qquad (3.12)$$

where,

$$p_{N_i}(j) = \binom{N_i}{j} \tau^j (1-\tau)^{(N_i - j)},$$

$$C_2(j) = \sum\nolimits_{l=0}^{2} (-1)^{l+1} \binom{2}{l} p_{N_i - 2}(j - l),$$

$$V_2 = \frac{2\kappa}{N_i} \tau(1-\tau),$$

$$k = \frac{N_i \left\{ \left(N_i - 1 \right) \mathrm{rel}_{20y} \sigma_Y^2 - N_i \sigma_Y^2 + \mu_X \left[N_i - \mu_Y \right] \right\}}{2 \left\{ \mu_Y \left(N_i - \mu_Y \right) - \mathrm{rel}_{20y} \sigma_Y^2 \right\}}$$

and $\mathrm{rel}_{20\mathrm{y}} = \dfrac{N_i}{N_i - 1} \left(1 - \dfrac{\sum_i \mu\left(Y_i\right)\left[1 - \mu\left(Y_i\right)\right]}{\sigma_Y^2} \right).$

Table 3.2: Descriptive Statistics of the SAT Scale Scores

Score Level	Score	# of Items	Alpha Reliability	RMS (CSEM)	Scale Score Reliability from RMS (CSEM)	Mean	SD	Skew	Kurt	Min Possible (Observed)	Max Possible (Observed)
Total	Total	154	—	40	0.96	1000.3	192.5	0.32	–0.5	400 (550)	1600 (1540)
Section	Math	58	—	31	0.9	500.2	99.9	0.36	–0.1	200 (240)	800 (800)
Section	ERW	96	—	26	0.94	500.1	104.2	0.32	–0.6	200 (250)	800 (780)
Test	Reading	52	0.89	1.8	0.89	25	5.4	0.43	–0.3	10 (12)	40 (39)
Test	Writing and Language	44	0.89	1.8	0.89	25	5.5	0.14	–0.6	10 (10)	40 (40)
Test	Math	58	0.9	1.6	0.9	25	5	0.36	–0.1	10 (12)	40 (40)
Cross-Test	Analysis in Science	35	0.87	1.9	0.86	25	5	0.33	–0.4	10 (12)	40 (40)
Cross-Test	Analysis in History/ Social Studies	35	0.83	2.2	0.83	24.9	5.5	0.39	–0.5	10 (11)	40 (40)

From the compound binomial distribution of the observed scores, *CSEM*s at τ values of interest can be estimated for each true score corresponding to an observed score of interest (i.e., τ values such that $N_i\tau = Y$) where the error estimation involves the entire range of observed scores, (i.e., Y_j for $j = 0$ to N_i),

$$CSEM_{N_i\tau = Y, CompoundBinomial} = \sqrt{\sum_{j=0}^{N_i} \left\{Y_j - \left[\sum_{j=0}^{N_i} Y_j \hat{\Pr}\left(Y_j|\tau\right)\right]\right\}^2 \hat{\Pr}\left(Y_j|\tau\right)} \tag{3.13}$$

The compound binomial distribution often involves fitting a four parameter beta distribution for the proportion correct true scores, which can suggest a distribution of proportion correct true scores that is defined for a narrower range than $\tau = 0$ to 1 (i.e., the four parameter beta distribution, Kolen & Brennan, 2014; Lord, 1969).

The estimated *CSEM*s for scale scores can be obtained by replacing the Y_js with unrounded and rounded scale scores. The plots of the estimated scale score *CSEM*s for Math, Reading, Writing and Language, Analysis in Science, and Analysis in History/Social Studies are presented in Figures 3.1–3.5. The vertical lines in these figures denote parts of the proportion true score range that are defined and undefined in the estimation of the proportion correct distribution. *CSEM*s for all defined and undefined τ values were calculated using Equation 3.13. As shown in Figures 3.1–3.5, the *CSEM*s for all test and cross-test scores were very close to constant across all scores. Many of the fluctuations in the *CSEM*s shown in the figures can be attributed to rounding effects (i.e., the *CSEM* series based on rounded scale scores fluctuate more than the *CSEM* series based on unrounded scale scores).

Figure 3.1: *CSEM*s of the Adjusted, Rounded and Unrounded Scale Scores for SAT Reading (52 Items)

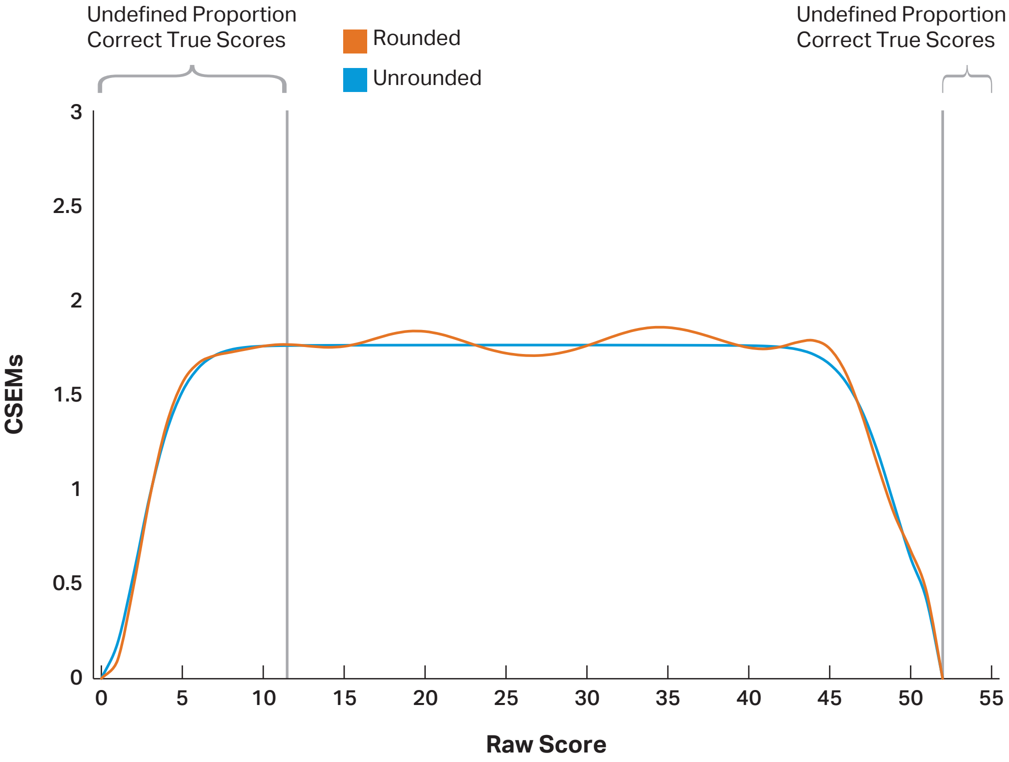


Figure 3.2: *CSEMs* of the Adjusted, Rounded and Unrounded Scale Scores for SAT Math (58 Items)

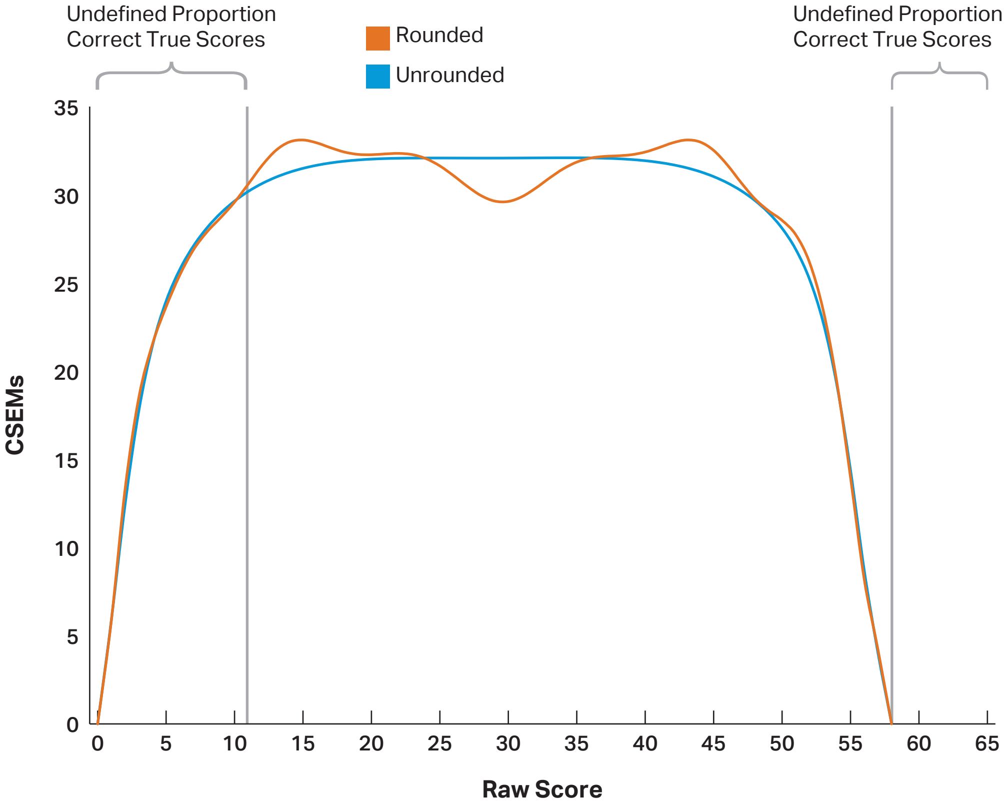


Figure 3.3: *CSEMs* of the Adjusted, Rounded and Unrounded Scale Scores for SAT Writing and Language (44 Items)

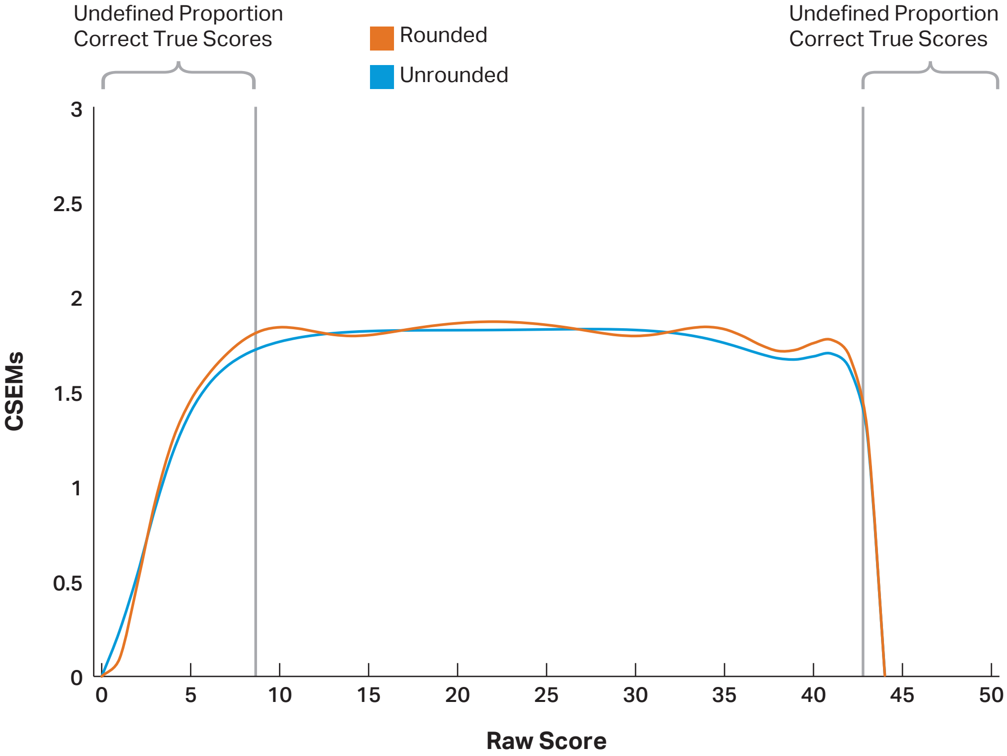


Figure 3.4: ***CSEM*s of the Adjusted, Rounded and Unrounded Scale Scores for SAT Analysis in Science (35 Items)**

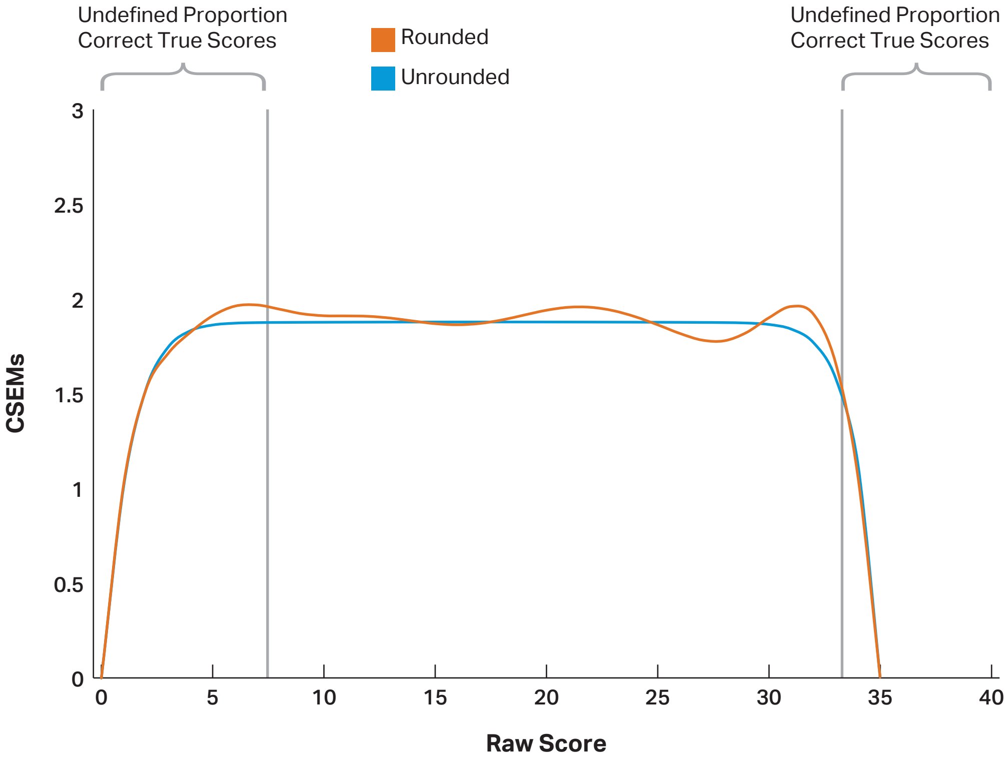
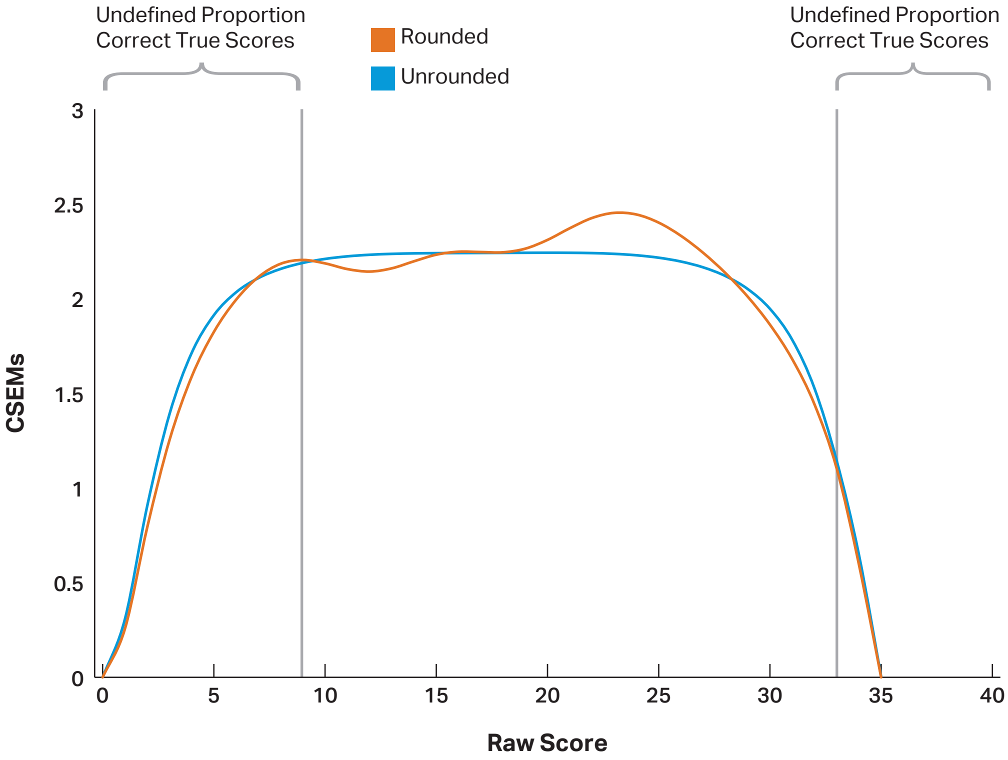


Figure 3.5: ***CSEM*s of the Adjusted, Rounded and Unrounded Scale Scores for SAT Analysis in History/Social Studies (35 Items)**



Estimation equations for the reliabilities of each scale score were developed from the *CSEM*s to obtain the scaling results reported in this chapter:

$$Reliability = 1 - \frac{MS(CSEM)}{\sigma_{SC}^{2}}, \tag{3.14}$$

where $\sigma_{sc}{}^{2}$ is the variance of scale scores. Cronbach's alpha was reported for raw score reliability. The mean squared *CSEM*, *MS*(*CSEM*), was obtained as the weighted average of the squared scale score *CSEM*s for all raw scores, for the scales directly established. Thus, the MS(*CSEM*) can be written as

$$MS(CSEM) \approx \int CSEM_{sc,\tau}^{\,2} Prob(\tau) d\tau, \tag{3.15}$$

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

For the scores that were mathematically derived including the Math Test score (Equation 3.10), the ERW section score (Equation 3.9), and total scores (Equation 3.11), the following equations were used to compute the root mean squared *CSEM*, *RMS*(*CSEM*):

$$RMS(CSEM)_{MTS} = \sqrt{\frac{MS(CSEM)_{MSS}}{20^{2}}} \tag{3.16}$$

$$RMS(CSEM)_{ERW} = \sqrt{MS(CSEM)_{R} \cdot 10^{2} + MS(CSEM)_{WL} \cdot 10^{2}} \tag{3.17}$$

$$RMS(CSEM)_{Total} = \sqrt{MS(CSEM)_{ERW} + MS(CSEM)_{MSS}}\,. \tag{3.18}$$

The estimated scale score *CSEM*s for test and cross-test scores range between 1.6 and 2.2. The estimated scale score *CSEM*s for total, Math, and ERW scores were 40, 31, and 26, respectively. Based on the methods described above, the estimated scale score reliabilities for Reading, Writing and Language, Math, Analysis in Science, and Analysis in History/Social Studies scores were 0.89, 0.89, 0.90, 0.86, and 0.83, respectively. The estimated scale score reliabilities for total, Math, and ERW scores were 0.96, 0.90, and 0.94, respectively (Table 3.2).

Evaluation 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 alternate forms. The scales have also been evaluated in terms of meeting the scaling goals with respect to *CSEM*s, score gaps, many-to-one conversions, minimum and maximum possible scores, etc.

The Standards call for warning test users of the limitations and potential misinterpretations of the reporting scales (Standards 5.1 and 5.3, AERA/APA/NCME, 2014). A particularly important limitation and potential misinterpretation of the SAT scales is that the scales were not established based on the performance of operational SAT examinees. 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 where demographic characteristics were approximated through weighting. The screenings and weightings were selected from several plausible options. Because of the limitations of the SAT scales, 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. 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 (e.g., Standard 5.6, AERA, APA, & NCME, 2014, p. 103). The scales established for the redesigned SAT should be continually evaluated for indications that revisions are warranted.

BIBLIOGRAPHY/REFERENCES

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

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics, 12,* 971–988.

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (Research Memorandum No.14-01). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*(4), 285–307.

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika, 30,* 239–270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika, 34,* 259–299.

Moses, T., & Golub-Smith, M. (2011). *A scaling method that produces scale score distributions with specific skewness and kurtosis* (ETS Research Memorandum, ETS RM-11-04). Princeton, NJ: Educational Testing Service.

Moses, T., & Kim, Y. K. (2017). Stabilizing conditional errors of measurement in scale score transformation. *Journal of Educational Measurement, 54*(2), 184–199.

CHAPTER 4

Characteristics of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 Vertical Scalings

Tim Moses and YoungKoung Kim

Goals for the Vertical Scales

The vertical scales for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 tests were established using non-operational data obtained from the 2014 scaling study (i.e., the same study used for the SAT scaling). The PSAT/NMSQT and PSAT 10 and PSAT 8/9 scales were developed to support a vertically aligned longitudinal assessment system that can support evaluations of student growth across:

- Subject domain (Reading, Math, Writing and Language, Analysis in History/Social Studies, and Analysis in Science);
- Nationally representative test performance in high school grades (i.e., 9th graders taking the PSAT 8/9, 10th graders taking the PSAT/NMSQT and PSAT 10, and 11th graders taking the SAT);
- Testing programs in the SAT Suite (SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9).

The vertical scales for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 were established by linking number correct scores on base forms of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 to the established SAT scales (described in Chapter 3) using a scaling test design (described in Chapters 1 and 2). By linking the number correct scores on the base forms of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 to the SAT scales, the PSAT/NMSQT and PSAT 10 and PSAT 8/9 scales were vertically linked to, and generally reflective of, the characteristics of the SAT scales described in Chapter 3, except for the following features:

- The Reading Test scores, Writing and Language Test scores, the Analysis in Science cross-test scale scores, and Analysis in History/Social Studies cross-test scale scores were set with ranges of 8–38 for the PSAT/NMSQT and PSAT 10, and 6–36 for the PSAT 8/9;
- The Math and Evidence-Based Reading and Writing (ERW) section scores were set with ranges of 160–760 for the PSAT/NMSQT and PSAT 10, and 120–720 for the PSAT 8/9;
- All correct, maximum possible raw scores convert to the highest obtainable PSAT/NMSQT and PSAT 10 and PSAT 8/9 scale scores;
- None correct, minimum possible raw scores convert to the lowest obtainable PSAT/NMSQT and PSAT 10 and PSAT 8/9 scale scores.

Unlike the procedures used for SAT scaling, the scaling procedures for the PSAT/NMSQT and PSAT 10 and the PSAT 8/9 allowed for less control of the gaps, many-to-one conversions, *CSEM*s stability, means, and other aspects of the scale score distributions.

Method

Data. The PSAT/NMSQT and PSAT 10 and PSAT 8/9 scales are intended to reflect nationally representative test performance under motivated test conditions. To obtain the desired examinee samples for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 scaling, students were recruited from a national distribution of high schools to voluntarily take one of three SAT test forms, one of three PSAT/NMSQT and PSAT 10 test forms, or one PSAT 8/9 test form in December 2014 (Tables 2.3 and 2.4). The assumption was that the high schools College Board recruited to participate in the vertical scaling study would produce student samples that would convey nationally representative performance. The form selected as the base form for the SAT scaling (see Chapter 3) was also used as the base form for the vertical scaling study. The base form for the PSAT/NMSQT and PSAT 10 was determined based on an evaluation of the PSAT/NMSQT and PSAT 10 forms' statistical properties. The resulting scaling samples were made up of 6,024 11th- and 12th-grade students taking the SAT form; 6,443 10th- and 11th-grade students taking the PSAT/NMSQT and PSAT 10 form; and 11,014 9th-grade students taking the PSAT 8/9 form. The forms were complete SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 tests containing the Reading, Math, and Writing and Language Tests, which included the Analysis in Science and Analysis in History/Social Studies items, and the items for the seven subscores (PSAT/NMSQT and PSAT 10) and six subscores (PSAT 8/9) that constitute the redesigned assessments.

Because the vertical scaling was based on a scaling test design, students also completed one of five scaling tests with items corresponding to the Reading Test, the Math Test, the Writing and Language Test, Analysis in Science, or Analysis in History/Social Studies. These scaling tests were designed to represent the content and difficulty levels of the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 tests for one specific subject domain, and to also be administered as separately timed, 60-minute tests that are external to their corresponding test. The number of items in each scaling test is as follows:

- 44 items on the scaling test for Reading;
- 45 items on the scaling test for Math;
- 72 items on the scaling test for Writing and Language;
- 51 items on the scaling test for Analysis in Science;
- 51 items on the scaling test for Analysis in History/Social Studies.

As shown in Table 4.1, the same scaling tests were randomly administered to five similarly sized subgroups of the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 student samples. That is, one-fifth of the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 tests were packaged with the Reading scaling test; spiraled with tests packaged with one of the other scaling tests; and administered to approximately one-fifth of the SAT examinees, the PSAT/NMSQT and PSAT 10 examinees, and the PSAT 8/9 examinees. The administration of the scaling test and either the SAT, PSAT/NMSQT and PSAT 10, or PSAT 8/9 tests to the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 examinees is depicted in Table 4.1, as it was applied to the subsamples taking the Reading, Math, Writing and Language, Analysis in Science, and Analysis in History/Social Studies scaling tests. All of the data evaluations, screenings, weightings, and scalings were implemented on the subsamples who took the SAT, PSAT/NMSQT and PSAT 10 and PSAT 8/9 (i.e., the five subsamples who took the Reading, Math, Writing and Language, Analysis in Science, and Analysis in History/Social Studies scaling tests).

Table 4.1: Depiction of the Scaling Test Design for the Vertical Scalings

	SAT	PSAT/NMSQT and PSAT 10	PSAT 8/9
Scaling Test	√	√	√
SAT Test	√		
PSAT 10 Test		√	
PSAT 8/9 Test			√

For motivation screening, motivated students were identified based on test performance and also on their response to a survey question about their effort given on each assessment,[1] as:

- A maximum percentage of either 25% or 50% omitted items on the Reading, Math, Writing and Language, and scaling tests;
- An absolute difference in standardized scores on the scaling test and on the corresponding test or cross-test items less than 3;
- Answering at least one student-produced response (SPR) item on the operational Math Test (not the scaling test);
- A response to the survey question of effort indicating that they either tried their best, began by trying their best, or gave moderate effort.

Grade-level participation on the SAT and PSAT/NMSQT and PSAT 10 was also considered in terms of motivation and target populations for the scaling. The combination of screenings for the samples resulted in four potential screenings of the samples for the SAT and PSAT/NMSQT and PSAT 10 forms and two potential samples for the PSAT 8/9 form (Table 4.2). After evaluating the potential samples with respect to sample size, average test performance, and desired representation of the targeted populations, the samples selected from the motivation screenings were Screening 3. This means that examinees did not indicate giving little effort on the test; completed at least one SPR Math item; had standard scores on the scaling test that differed by less than +/−3 with the test or cross-test items corresponding to that scaling test; omitted no more than 50% of items on the Reading, Math, Writing and Language, and scaling tests; and were 11th-, 10th-, and 9th-grade students taking the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 forms, respectively.

The students taking the SAT who were used for the vertical scaling study were different from the students taking the SAT who were used to set the SAT scales (see Chapter 3). The vertical scaling study sample for the SAT was intended to be a nationally representative group of 11th graders, whereas the SAT scaling study sample was supposed to reflect the characteristics of the college-bound SAT cohort group of 11th and 12th graders. To make a distinction between the two groups, the national group is referred to as the NRSAT group while the SAT scaling group based on weighting to the historical SAT cohort is referred to as the “SAT (Cohort)” group in this chapter.

[1] Survey Question 1, shown in the Appendix to Chapter 2, asked students to “Rate the level of effort that you gave while completing this test.” The response options were “1 = I tried my best”; “2 = I gave moderate effort”; “3 = I gave little effort”; “4 = I began by trying my best, but then I found the test very difficult; by the end of the test, I was no longer putting in much effort.”

Table 4.2: Sample Sizes of the Screened Samples Reviewed for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 Vertical Scalings

	NRSAT				PSAT/NMSQT and PSAT 10 Group				PSAT 8/9 Group	
	Screening 1	Screening 2	Screening 3*	Screening 4	Screening 1	Screening 2	Screening 3*	Screening 4	Screening 1	Screening 3*
%Omit Criterion	25%	25%	50%	50%	25%	25%	50%	50%	25%	50%
Grades	11th only	11th + 12th	11th only	11th + 12th	10th only	10th + 11th	10th only	10th + 12th	9th only	9th only
	Total Sample Sizes									
	2869	4282	3411	5023	3424	4334	4136	5164	7724	8741
	Scaling Test & Subsample Sample Sizes									
Reading	608	917	713	1060	726	930	850	1078	1626	1799
Math	571	862	695	1026	690	843	850	1033	1480	1697
Writing	579	870	689	1023	701	892	863	1074	1588	1804
Science	570	825	673	970	682	858	825	1020	1566	1758
History	541	808	641	944	625	811	748	959	1464	1683

* Denotes the screening and samples selected for the vertical scalings.

The screened NRSAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 student samples were evaluated for national representativeness by comparing distributions of students' survey responses to those from national high school surveys reported in recent publications from the National Center for Education Statistics (see Chapter 2). The variables of most interest in this evaluation were distributions of student subgroups from private and public high schools, gender, ethnicity for private and public school students, College Board region for private and public school students, high school urbanicity (i.e., students from city, rural, town, or suburban high schools) for private and public school students, and self-reported interest to pursue postsecondary education based on the survey. The student samples were weighted so that their distributions on the variables of interest would approximate the percentages from nationally surveyed high school students in reports available at https://nces.ed.gov/. Weighted and unweighted vertical scaling data for the five subsamples of students taking one of the five scaling tests for SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 are described in Tables 4.3–4.10. In these tables, private and public school student percentages are separated by the "/" symbol.

Some aspects of the vertical scaling data made the weighting approximation less precise than desired, such as sample sizes that were small overall (less than 1,000 for the NRSAT, PSAT/NMSQT and PSAT 10 samples), and zero for one or more subgroups of interest. The most difficult variable to approximate target proportions was urbanicity (i.e., students from rural, suburban, town, or urban schools). For urbanicity, recruitment and data gathering did not include a separate category for "town" that addressed nationally representative targets. Weighting for urbanicity was considered with respect to two options: one where town and rural school proportions were combined and another where the town proportions were distributed across rural, suburban, and urban proportions. Although the first of these options was closer to how urbanicity was treated in the school recruitment, evaluations of town designations for schools in the scaling study data revealed that (1) town designations for high schools can change with newer versus older data sources; and that (2) in the data, some high schools classified as rural, suburban, or town could have plausibly received town designations based on different data sources. Based on these evaluations and also on comparisons of scaling results from alternative weighting implementations, weighting by urbanicity was implemented by distributing the targeted town proportions across rural, suburban, and urban school students. This resulted in weighted datasets that were more closely reflective of the town designations in the actual data, and also not extremely different (within 3%–4%) from the original national targets for rural, suburban, and urban school students.

Scaling Methods and Procedures. Methods for vertical scaling were used to obtain conversions for the number-correct scores of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 test forms to the raw and scale scores of the SAT that:

- Account for the difficulty differences of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 test forms as compared to the SAT form;
- Preserve the ability differences of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 groups as compared to the NRSAT group;
- Reflect the PSAT/NMSQT and PSAT 10 and PSAT 8/9 examinee ability differences, but not the PSAT/NMSQT and PSAT 10 and PSAT 8/9 test form difficulty differences on a common scale (i.e., the scales developed for the SAT in the SAT scaling study).

Table 4.3: Samples Reviewed for NRSAT (unweighted)

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	92	91	90	90	90
Female (vs. Male)	50	51	54	53	53	52
Private/Public	5/95	4/96	4/96	4/96	5/95	5/95
	Private/Public					
American Indian	<1/1	./1	./1	./1	<1/1	./1
Asian	<1/4	./8	<1/12	<1/11	<1/8	<1/9
Black	<1/13	1/14	<1/15	<1/15	<1/16	<1/17
Hispanic	1/19	1/25	1/24	1/25	1/24	1/23
White	3/55	3/39	3/40	3/38	3/40	3/41
Other or Missing	<1/2	./	./	./	./	./.
MRO	1/21	./11	./11	./10	./10	./10
MSRO	1/14	./5	./5	./5	./5	./4
NERO	<1/4	./<1	./<1	./<1	./<1	./<1
SRO	1/21	1/19	1/21	1/22	<1/21	<1/23
SWRO	<1/11	1/42	1/40	1/40	1/42	1/40
WRO	1/23	3/19	3/18	3/18	4/17	3/17
Rural	<1/26	2/37	2/39	2/36	2/38	2/38
Suburban	2/31	2/31	2/33	2/33	2/33	2/33
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	1/27	1/23	1/26	1/25	1/25

Table 4.4: Samples Reviewed for NRSAT (weighted)

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	72	72	72	72	72
Female (vs. Male)	50	50	50	50	50	50
Private/Public	5/95	6/94	6/94	6/94	6/94	6/94
	Private/Public					
American Indian	<1/1	./.	./<1	./1	<1/1	./1
Asian	<1/4	./4	1/4	1/4	<1/4	<1/4
Black	<1/13	1/13	1/13	1/13	<1/13	<1/13
Hispanic	1/19	<1/19	1/19	1/19	1/19	1/19
White	3/55	4/55	4/55	4/55	4/55	4/55
Other or Missing	<1/2	./	./	./.	./	./.
MRO	1/21	./21	./21	./21	./21	./21
MSRO	1/14	./14	./14	./14	./14	./14
NERO	<1/4	./4	./4	./4	./4	./<4
SRO	1/21	2/21	2/21	2/21	2/21	2/21
SWRO	<1/11	1/11	2/11	2/11	2/11	2/11
WRO	1/23	3/22	2/22	2/22	2/22	2/22
Rural	<1/26	1/29	1/29	1/29	1/29	1/29
Suburban	2/31	2/34	2/34	2/34	2/34	2/34
Town	<1/11	./.		./.	./.	./.
Urban	2/27	3/30	3/30	3/30	3/30	3/30

Table 4.5: Samples Reviewed for the PSAT/NMSQT and PSAT 10 Vertical Scaling (unweighted)

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	86	88	86	89	84
Female (vs. Male)	50	53	53	53	55	57
Private/Public	5/95	5/95	6/94	5/95	4/96	5/95
	Private/Public					
American Indian	<1/1	<1/2	./1	./1	<1/1	./1
Asian	<1/4	<1/11	1/11	<1/13	<1/13	<1/13
Black	<1/13	<1/13	1/11	1/13	<1/11	<1/13
Hispanic	1/19	1/24	1/25	1/26	1/26	1/24
White	3/55	3/38	4/38	3/34	3/39	3/38
Other or Missing	<1/2	./.	./.	./.	./.	./.
MRO	1/21	./7	./6	./6	./7	./6
MSRO	1/14	./7	./6	./8	./8	./8
NERO	<1/4	./3	./2	./2	./3	./2
SRO	1/21	2/25	2/27	2/24	1/24	2/25
SWRO	<1/11	1/26	1/25	<1/24	<1/25	1/24
WRO	1/23	3/28	3/27	3/30	3/29	3/30
Rural	<1/26	1/35	1/36	1/35	1/36	1/36
Suburban	2/31	3/38	3/38	3/36	3/38	3/35
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	1/22	1/20	1/24	1/22	1/24

Table 4.6: Samples Reviewed for the PSAT/NMSQT and PSAT 10 Vertical Scaling (weighted)

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	72	72	72	72	72
Female (vs. Male)	50	50	50	50	50	50
Private/Public	5/95	6/94	6/94	6/94	6/94	6/94
	Private/Public					
American Indian	<1/1	<1/<1	./<1	./1	<1/<1	./<1
Asian	<1/4	1/4	<1/4	<1/4	<1/4	1/4
Black	<1/13	1/13	1/13	1/13	<1/13	1/13
Hispanic	1/19	1/19	1/19	1/19	1/19	1/19
White	3/55	4/55	4/55	4/55	5/55	4/55
Other or Missing	<1/2	./.	./.	./.	./.	./.
MRO	1/21	./21	./21	./21	./21	./21
MSRO	1/14	./14	./14	./14	./14	./14
NERO	<1/4	./4	./4	./4	./4	./4
SRO	1/21	2/21	2/21	2/21	2/21	2/21
SWRO	<1/11	2/11	2/11	2/11	2/11	2/11
WRO	1/23	2/22	2/22	2/22	2/23	2/22
Rural	<1/26	1/29	1/29	1/29	1/29	1/29
Suburban	2/31	2/34	2/34	2/34	2/35	2/34
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	3/30	3/30	3/30	3/31	3/30

Table 4.7: Samples Reviewed for the PSAT/NMSQT and PSAT 10 Vertical Scaling (unweighted)—Second Form

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	89	89	88	88	88
Female (vs. Male)	50	52	53	50	52	52
Private/Public	5/95	6/94	5/95	5/95	6/94	4/96
	Private/Public					
American Indian	<1/1	./1	<1/1	<1/<1	<1/1	./1
Asian	<1/4	<1/11	<1/11	<1/12	<1/12	<1/11
Black	<1/13	1/12	1/14	<1/11	1/12	./12
Hispanic	1/19	1/27	1/24	1/24	1/26	1/29
White	3/55	4/36	3/39	3/41	4/38	3/37
Other or Missing	<1/2	./.	./.	./.	./.	./.
MRO	1/21	./6	./7	./7	./6	./7
MSRO	1/14	./8	./7	./8	./7	./8
NERO	<1/4	./2	./2	./2	./2	./3
SRO	1/21	2/25	2/27	2/24	3/26	1/23
SWRO	<1/11	1/25	<1/24	1/25	<1/24	1/26
WRO	1/23	3/29	3/28	3/28	3/28	3/30
Rural	<1/26	1/37	1/36	1/36	1/36	1/35
Suburban	2/31	3/35	3/36	3/37	3/36	3/38
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	2/22	1/23	1/22	1/22	<1/23

Table 4.8: Samples Reviewed for the PSAT/NMSQT and PSAT 10 Vertical Scaling (weighted)—Second Form

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	72	72	72	72	72
Female (vs. Male)	50	50	50	50	50	50
Private/Public	5/95	6/94	6/94	6/94	6/94	6/94
	Private/Public					
American Indian	<1/1	./1	<1/<1	<1/<1	<1/<1	./<1
Asian	<1/4	<1/4	<1/4	<1/4	<1/4	1/4
Black	<1/13	1/13	1/13	<1/13	1/13	./13
Hispanic	1/19	1/19	1/19	1/19	1/19	1/19
White	3/55	4/55	4/55	4/55	4/55	4/55
Other or Missing	<1/2	./.	./.	./.	./.	./.
MRO	1/21	./21	./21	./21	./21	./21
MSRO	1/14	./14	./14	./14	./14	./14
NERO	<1/4	./4	./4	./4	./4	./4
SRO	1/21	2/21	2/21	2/21	2/21	2/21
SWRO	<1/11	2/11	2/11	2/11	2/11	1/11
WRO	1/23	2/22	2/22	2/22	2/22	2/23
Rural	<1/26	1/29	1/29	1/29	1/29	1/29
Suburban	2/31	2/34	2/34	2/34	2/34	2/34
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	3/30	3/30	3/30	3/30	3/31

Table 4.9: Samples Reviewed for the PSAT 8/9 Vertical Scaling (unweighted)

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	84	84	85	85	83
Female (vs. Male)	50	51	53	51	51	49
Private/Public	5/95	8/92	8/92	8/92	8/92	8/92
	Private/Public					
American Indian	<1/1	<1/1	./1	./1	<1/1	<1/1
Asian	<1/4	<1/10	<1/12	<1/11	<1/9	<1/9
Black	<1/13	1/12	1/10	1/12	1/12	1/12
Hispanic	1/19	1/21	1/20	1/22	1/22	1/22
White	3/55	5/41	5/41	6/38	6/41	5/41
Other or Missing	<1/2	./.	./.	./.	./.	./.
MRO	1/21	./13	./15	./13	./14	./14
MSRO	1/14	./4	./5	./4	./4	./4
NERO	<1/4	./3	./2	./2	./2	./2
SRO	1/21	3/30	3/29	3/30	3/31	3/32
SWRO	<1/11	1/18	1/15	1/18	1/17	1/17
WRO	1/23	4/23	4/26	4/24	4/23	4/24
Rural	<1/26	2/33	3/34	2/34	3/34	2/35
Suburban	2/31	4/35	4/34	4/33	4/34	4/34
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	2/23	2/24	2/24	1/24	2/23

Table 4.10: Samples Reviewed for the PSAT 8/9 Vertical Scaling (weighted)

Subgroup	Nationally Representative	Reading	Math	Writing	Science	History
Postsecondary Intention	72	72	72	72	72	72
Female (vs. Male)	50	50	50	50	50	50
Private/Public	5/95	6/94	6/94	6/94	6/94	6/94
	Private/Public					
American Indian	<1/1	<1/1	./1	./1	<1/<1	<1/<1
Asian	<1/4	<1/4	<1/4	<1/4	<1/4	1/4
Black	<1/13	<1/13	<113	<1/13	<1/13	<1/13
Hispanic	1/19	<1/19	<1/19	1/19	1/19	1/19
White	3/55	4/55	4/55	4/55	4/55	4/55
Other or Missing	<1/2	./.	./.	./.	./.	./.
MRO	1/21	./21	./21	./21	./21	./21
MSRO	1/14	./14	./14	./14	./14	./14
NERO	<1/4	./4	./4	./4	./4	./4
SRO	1/21	3/21	3/21	2/21	2/21	2/21
SWRO	<1/11	2/11	2/11	2/11	2/11	2/11
WRO	1/23	2/22	2/22	2/22	2/22	2/22
Rural	<1/26	1/29	1/29	1/29	1/29	1/29
Suburban	2/31	2/34	2/34	2/34	2/34	2/34
Town	<1/11	./.	./.	./.	./.	./.
Urban	2/27	3/30	3/30	3/30	3/30	3/30

The vertical scaling analyses utilized chained equipercentile conversions to express the raw scores of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 tests and cross-tests first on the scales of the raw scaling test forms for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 subsamples, and then to chain these converted scores from the scaling test percentiles to the raw scores of the corresponding SAT test and cross-tests for the NRSAT subsamples. The equipercentile conversion from the score *X* of one of the PSAT-related assessments, $PSAT_X$, to the scaling test can be expressed as,

$$e_{ScalingTest}\left(PSAT_X\right) = H^{-1}_{PSAT}\left[F\left(PSAT_X\right)\right] \tag{4.1}$$

where *F* and H_{PSAT} denote percentile rank functions based on the PSAT/NMSQT and PSAT 10 or PSAT 8/9 and scaling test score distributions obtained from the PSAT/NMSQT and PSAT 10 or PSAT 8/9 group.

The equipercentile conversion from score *S* of one of the scaling tests, $ScalingTest_S$, to the SAT test can be expressed as,

$$e_{NRSAT}\left(ScalingTest_s\right) = G^{-1}\left[H_{NRSAT}\left(ScalingTest_s\right)\right] \tag{4.2}$$

where *G* and H_{NRSAT} denote percentile rank functions based on the SAT and scaling test score distributions obtained from the NRSAT group.

The two equipercentile functions in Equations (4.1) and (4.2) are chained together as,

$$e_{NRSAT}\left[e_{ScalingTest}\left(PSAT_X\right)\right] = G^{-1}\left\langle H_{NRSAT}\left\{H^{-1}_{PSAT}\left[F\left(PSAT_X\right)\right]\right\}\right\rangle \tag{4.3}$$

and the resulting PSAT-to-Scaling Test-to-SAT score conversions are expressed on the SAT scale using interpolations of the raw-to-scale conversions for the SAT. Chaining the equipercentile conversions through the scaling test produced raw score conversions that preserved the ability differences of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 versus NRSAT groups observed on the scaling tests.

Additional procedures were used to increase the statistical stability of the chained equipercentile conversions obtained from the scaling test subsamples that were smaller than desirable (Table 4.2). Two smoothing procedures were considered, the first of which involved smoothing the score distributions used in the chained equipercentile conversions (loglinear presmoothing),

$$\log\left(m_X\right) = \sum\nolimits_{r=0}^{R} \beta_r X^r \tag{4.4}$$

where m_x is the expected frequency at score *X* of one of the four test score distributions involved in the chained equipercentile conversions. The *R* represents the number of parameters, *β*, that are estimated, where the estimation results in the first *R* moments of the observed score distribution being fit in the presmoothed distribution. A second smoothing method considered was cubic spline postsmoothing of the unsmoothed chained equipercentile conversions (Kolen & Brennan, 2014). Scaling results were reviewed based on different degrees of pre- and postsmoothing. The smoothing method ultimately used was the loglinear presmoothing method, fitting the first four, five, or six moments of the total test and scaling test distributions based on statistical tests, graphical inspections of the test score distributions, and characteristics of the resulting scale scores.

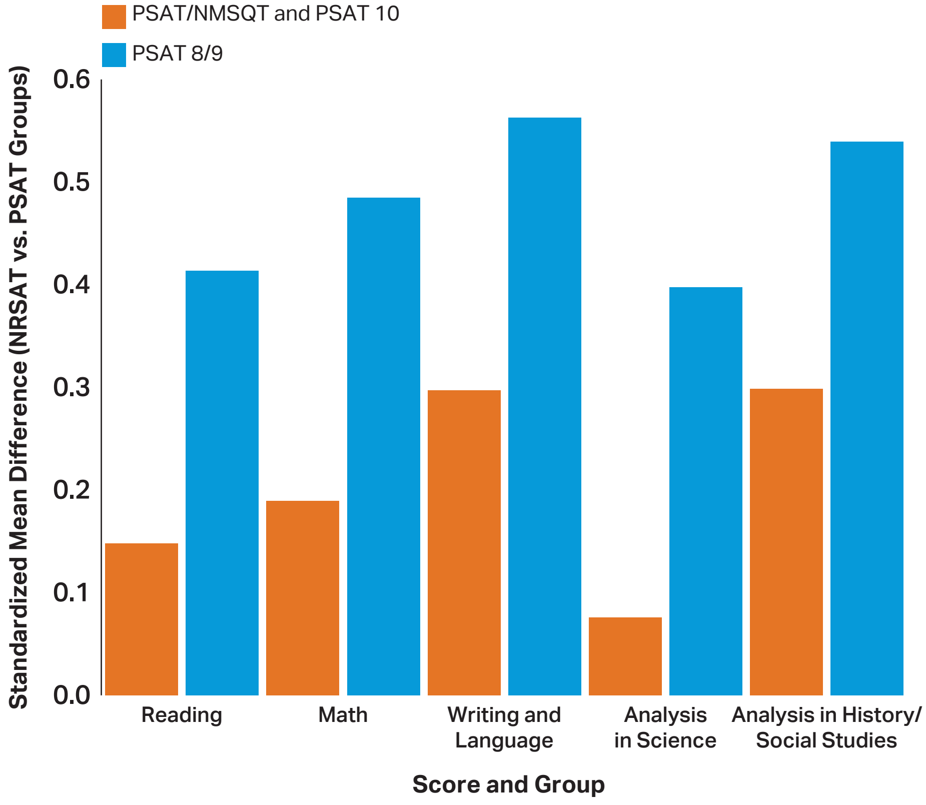
For the PSAT/NMSQT and PSAT 10 conversions, additional data were leveraged from one of the other two PSAT/NMSQT and PSAT 10 test forms administered in a spiraled administration with the base form (Tables 2.3 and 2.4). The data from students taking this second PSAT/NMSQT and PSAT 10 form were screened, weighted, smoothed, and scaled to SAT scales in the same way as the base form for PSAT/NMSQT and PSAT 10. The SAT scale score conversion for the PSAT/NMSQT and PSAT 10 base form was obtained as a weighted average of the base form conversion to the SAT scale described previously (weighted 0.7) and the conversion of the base form to the second PSAT/NMSQT and PSAT 10 form to the SAT scale (weighted 0.3).

For the final steps, the PSAT/NMSQT and PSAT 10 and PSAT 8/9 scores that were converted to the raw SAT form scales were expressed on the established SAT scales using interpolations of the raw-to-raw SAT chained equipercentile conversion tables and the raw SAT-to-scale conversion tables established in prior SAT scaling analyses. Linear interpolations (see Chapter 3, Equation 3.8) were applied to the lowest and highest PSAT-to-SAT scale scores with less than approximately 1% of examinees to ensure that when rounded, these scale scores would reflect the intended scale score ranges (i.e., ranges of 160–760 and 120–720 for the Math section score, and ranges of 8–38 and 6–36 for the other test and cross-test scores). The converted scale scores were then rounded to the intended scale score increments (i.e., an increment of 10 for the Math section score and an increment of 1 for the other non-Math test and cross-test scores).

Results

To assess the scaling goal to reflect the ability differences of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 students relative to the NRSAT students, the estimated ability differences were reviewed as standardized mean differences of the vertical scaling subsamples that completed the same scaling test (Figure 4.1). Figure 4.1 shows that the ability differences

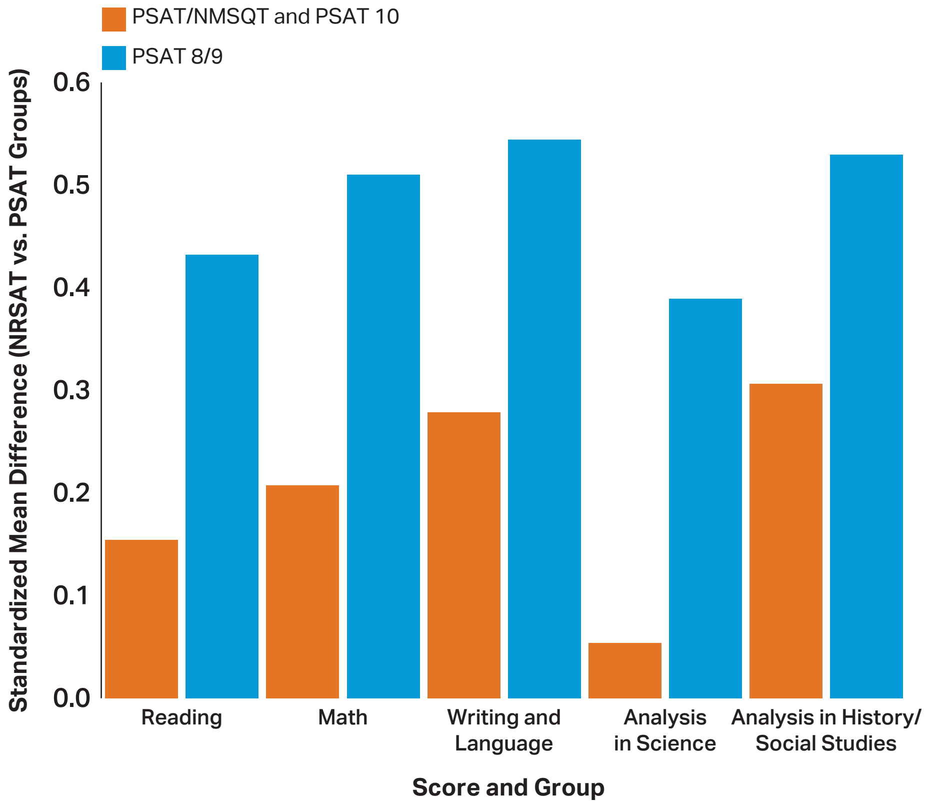
Figure 4.1: Standardized Mean Differences vs. NRSAT Raw Scaling Test Scores



of the NRSAT group to the PSAT/NMSQT and PSAT 10 group are estimated to range from approximately 0.07 (Analysis in Science) to about 0.30 (Writing and Language and Analysis in History/Social Studies) standard deviation units. Standardized mean differences for the NRSAT to the PSAT 8/9 group range from approximately 0.41 (Reading and Analysis in Science) to about 0.54 (Writing and Language and Analysis in History/Social Studies).

Vertical Scaling: Conversion Tables. The conversion tables for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 test forms were selected after several iterations of reviewing screened and weighted data, a range of pre- and post-smoothing degrees, different weighted averages of the conversions for the PSAT/NMSQT and PSAT 10 forms and examinee groups, and linear adjustments for different ranges of the highest and lowest scale scores. One way to assess the effectiveness of the vertical scaling procedures is to compare the standardized mean differences between the groups on the rounded and truncated scale scores (Figure 4.2) to those on the scaling test raw scores (Figure 4.1). Figure 4.2 shows that the standardized mean differences of the groups on the preliminary scale scores are similar to those observed on the raw scaling test in Figure 4.1. Similar to Figure 4.1, Figure 4.2 shows that standardized mean differences for the NRSAT and PSAT/NMSQT and PSAT 10 groups range from approximately 0.06 (Analysis in Science) to about 0.30 (Writing and Language, and Analysis in History/Social Studies). Standardized mean differences for the NRSAT and PSAT 8/9 groups range from approximately 0.42 (Reading and Analysis in Science) to about 0.53 (Writing and Language, and Analysis in History/Social Studies).

Figure 4.2: Standardized Mean Differences vs. NRSAT Rounded and Truncated Scale Scores



Vertical Scaling Implications. A major interest in establishing the PSAT/NMSQT and PSAT 10 and PSAT 8/9 vertical scales was to reflect nationally representative performance differences on the SAT scale. Expectations for the vertical scales were that the scale score distributions of the college-bound SAT cohort group of 11th- and 12th-grade examinees used to establish the SAT scales (SAT) should be higher than those of the NRSAT group, which should be higher than those of the PSAT/NMSQT and PSAT 10 group, which should be higher than those of the PSAT 8/9 group. To evaluate these expectations, the scale scores were summarized for the complete vertical scaling samples (not the scaling test subsamples) of SAT cohort (N = 4,346), NRSAT (N = 3,411), PSAT/NMSQT and PSAT 10 (N = 4,136), and PSAT 8/9 examinees (N = 8,741). Comparisons were also made with respect to the scale score distributions of the SAT cohort group. Results are shown in terms of summary statistics (Table 4.11) for the total, section, test, and cross-test scale scores. In general, Table 4.11 shows that means and standard deviations reflect expectations that they increase from the PSAT 8/9 to the SAT cohort samples.

Another evaluation of interest for understanding the implications of the vertical scales is the assessment of the conditional standard errors of measurement (*CSEM*s). Although not directly controlled or stabilized in the vertical scaling process, the extent to which the *CSEM*s of the vertical scales for PSAT/NMSQT and PSAT 10 and PSAT 8/9 were stable was a secondary interest and goal (see Chapter 1). The *CSEM*s are presented for the vertical scales established for PSAT/NMSQT and PSAT 10 (Figures 4.3–4.7) and PSAT 8/9 (Figures 4.8–4.12). Unlike the *CSEM* figures shown in Chapter 3 from the SAT scaling, the *CSEM*s shown in Figures 4.3–4.12 are not completely stable, as their stability was not directly controlled in the vertical scaling process.

Table 4.11: Summary Statistics of the SAT, NRSAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 Scale Scores from the Scaling Study Samples

Program Group	Score (Level)	Items	Rel.	Mean	SD	Skew	Kurt	Min Poss. (Obs.)	Max Poss. (Obs.)	25th Percentile	50th Percentile	75th Percentile
SAT (Cohort)	Total	154	—	1000.3	192.5	0.3	−0.5	400 (550)	1600 (1540)	870	1000	1130
NRSAT	(Total)	154	—	972.6	180	0.3	−0.3	400 (550)	1600 (1520)	840	960	1090
PSAT/NMSQT and PSAT 10		139	—	939.7	166.4	0.4	−0.3	320 (540)	1520 (1510)	810	920	1050
PSAT 8/9		120	—	881.7	149.8	0.3	−0.3	240 (440)	1440 (1400)	770	870	990
SAT (Cohort)	Math	58	—	500.2	99.9	0.4	−0.1	200 (240)	800 (800)	430	500	560
NRSAT	(Section)	58	—	485.9	95.9	0.4	0.2	200 (240)	800 (800)	420	480	540
PSAT/NMSQT and PSAT 10		48	—	470.9	85.9	0.4	0.1	160 (230)	760 (760)	410	470	530
PSAT 8/9		38	—	439.2	78.4	0.1	0.2	120 (180)	720 (720)	390	440	490
SAT (Cohort)	EBRW	96	—	500.1	104.2	0.3	−0.6	200 (250)	800 (780)	420	500	570
NRSAT	(Section)	96	—	486.7	95.8	0.3	−0.5	200 (240)	800 (780)	410	480	560
PSAT/NMSQT and PSAT 10		91	—	468.8	94.3	0.3	−0.6	160 (250)	760 (750)	390	460	540
PSAT 8/9		82	—	442.5	83.7	0.4	−0.5	120 (230)	720 (700)	380	430	500
SAT (Cohort)	Reading	52	0.89	25	5.4	0.4	−0.3	10 (12)	40 (39)	21	24	28
NRSAT	(Test)	52	0.87	24.3	4.9	0.4	−0.2	10 (11)	40 (39)	20	24	28
PSAT/NMSQT and PSAT 10		47	0.85	23.7	4.9	0.4	−0.6	8 (11)	38 (38)	19	23	27
PSAT 8/9		42	0.87	22.4	4.3	0.4	−0.3	6 (10)	36 (36)	19	22	25
SAT (Cohort)	Writing	44	0.9	25	5.5	0.1	−0.6	10 (10)	40 (40)	21	25	29
NRSAT	(Test)	44	0.88	24.4	5.1	0.2	−0.6	10 (11)	40 (40)	20	24	28

Table 4.11 continued on next page

Table 4.11 continued from previous page

Program Group	Score (Level)	Items	Rel.	Mean	SD	Skew	Kurt	Min Poss. (Obs.)	Max Poss. (Obs.)	25th Percentile	50th Percentile	75th Percentile
PSAT/NMSQT and PSAT 10		44	0.87	23.2	5.1	0.2	−0.6	8 (10)	38 (38)	19	23	27
PSAT 8/9		40	0.85	21.8	4.6	0.3	−0.5	6 (9)	36 (36)	19	21	25
SAT (Cohort)	Math	58	0.91	25	5	0.4	−0.1	10 (12)	40 (40)	21.5	25	28
NRSAT	(Test)	58	0.89	24.3	4.8	0.4	0.2	10 (12)	40 (40)	21	24	27
PSAT/NMSQT and PSAT 10		48	0.83	23.5	4.3	0.4	0.1	8 (11.5)	38 (38)	20.5	23.5	26.5
PSAT 8/9		38	0.84	22	3.9	0.1	0.2	6 (9)	36 (36)	19.5	22	24.5
SAT (Cohort)	Science	35	0.87	25	5	0.3	−0.4	10 (12)	40 (40)	21	24	29
NRSAT	(Cross-Test)	35	0.85	24.3	4.8	0.3	−0.4	10 (12)	40 (40)	21	24	28
PSAT/NMSQT and PSAT 10		32	0.81	23.6	4.7	0.5	−0.3	8 (10)	38 (38)	20	23	27
PSAT 8/9		29	0.82	22.4	4.3	0.5	−0.1	6 (8)	36 (36)	19	22	25
SAT (Cohort)	History	35	0.83	24.9	5.5	0.4	−0.5	10 (11)	40 (40)	20	25	29
NRSAT	(Cross-Test)	35	0.8	24.4	5.2	0.3	−0.3	10 (11)	40 (39)	20	24	28
PSAT/NMSQT and PSAT 10		32	0.79	23.4	4.8	0.3	20.5	8 (11)	38 (38)	20	23	27
PSAT 8/9		29	0.8	21.9	4.5	0.4	20.2	6 (6)	36 (36)	18	21	25

Note. The statistics in this table were obtained from groups with sample sizes of 4,346 for the SAT (Cohort), 3,411 for the NRSAT group, 4,136 for the PSAT/NMSQT and PSAT10 group, and 8,741 for the PSAT8/9 group.

Figure 4.3: ***CSEMs* of the Adjusted, Rounded and Unrounded Scale Scores for PSAT/NMSQT and PSAT 10 Reading (47 Items)**

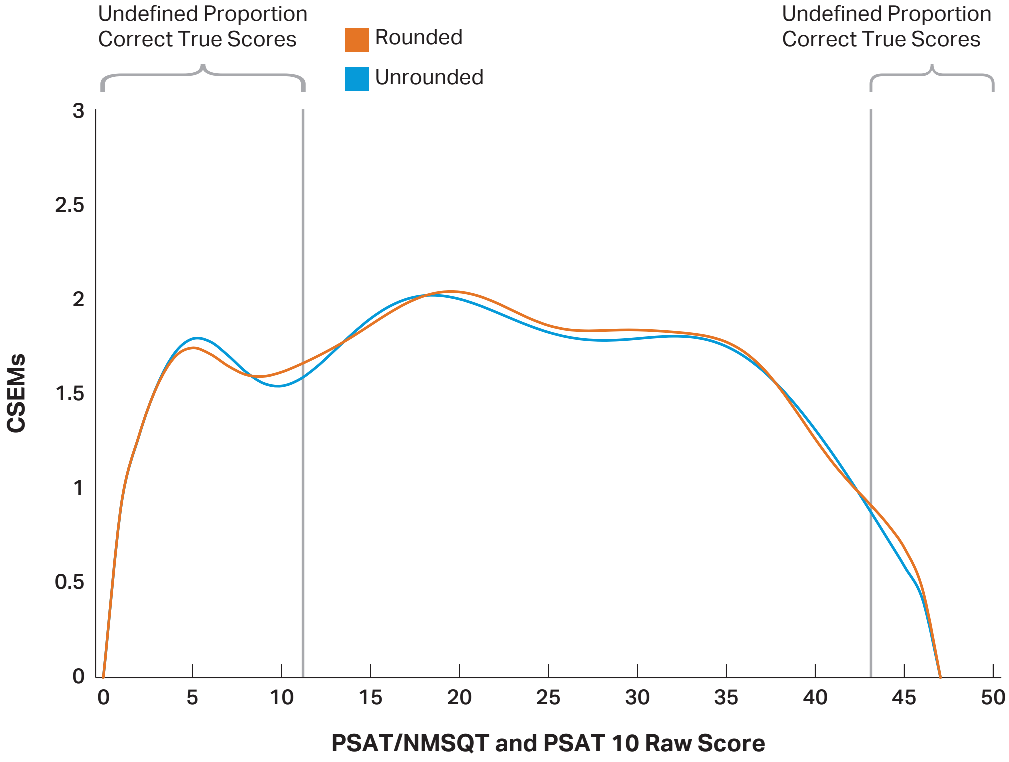


Figure 4.4: ***CSEMs* of the Adjusted, Rounded and Unrounded Scale Scores for PSAT/NMSQT and PSAT 10 Math (48 Items)**

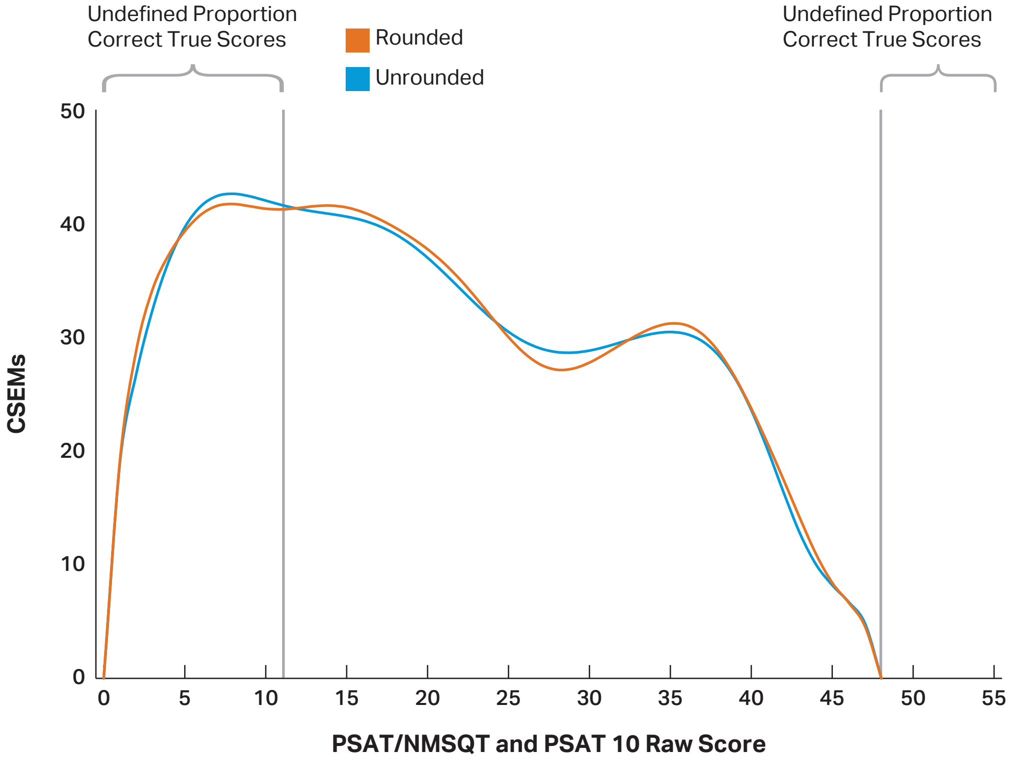


Figure 4.5: *CSEMs* of the Adjusted, Rounded and Unrounded Scale Scores for PSAT/NMSQT and PSAT 10 Writing and Language (44 Items)

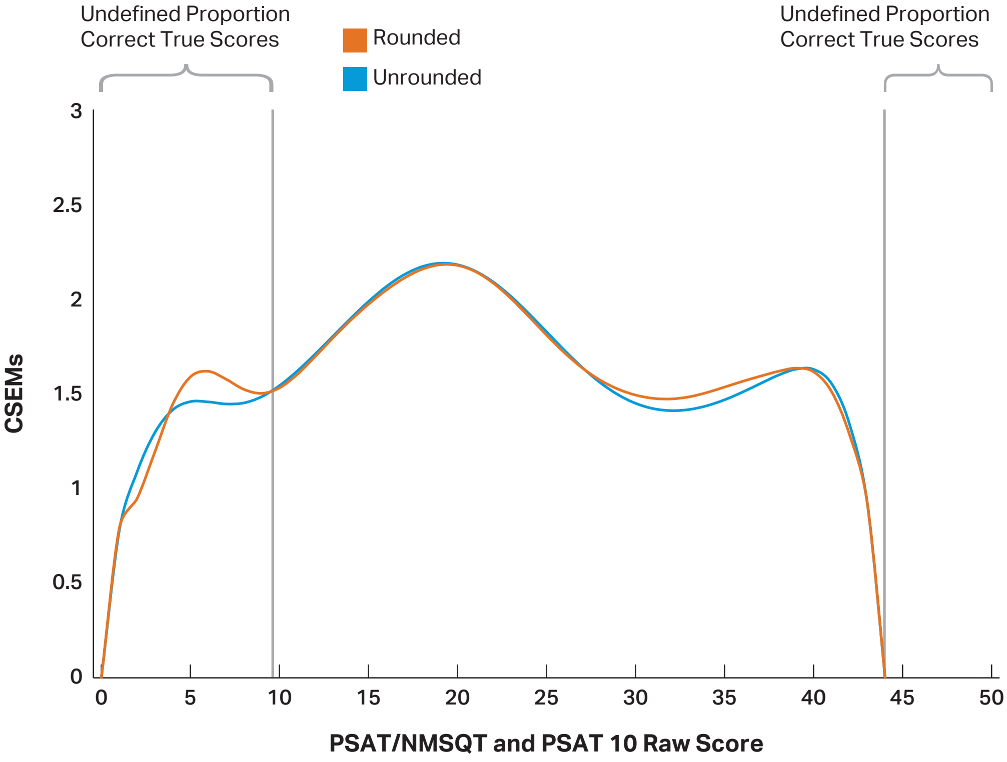


Figure 4.6: *CSEMs* of the Adjusted, Rounded and Unrounded Scale Scores for PSAT/NMSQT and PSAT 10 Analysis in Science (32 Items)

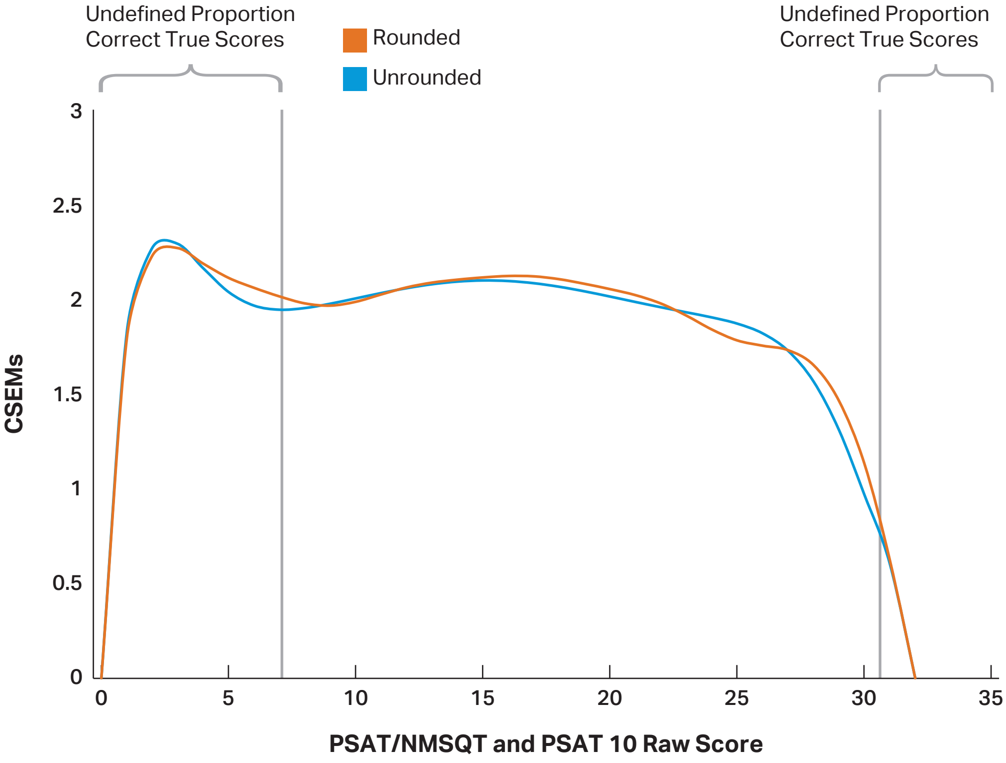


Figure 4.7: ***CSEMs*** **of the Adjusted, Rounded and Unrounded Scale Scores for PSAT/NMSQT and PSAT 10 Analysis in History/Social Studies (32 Items)**

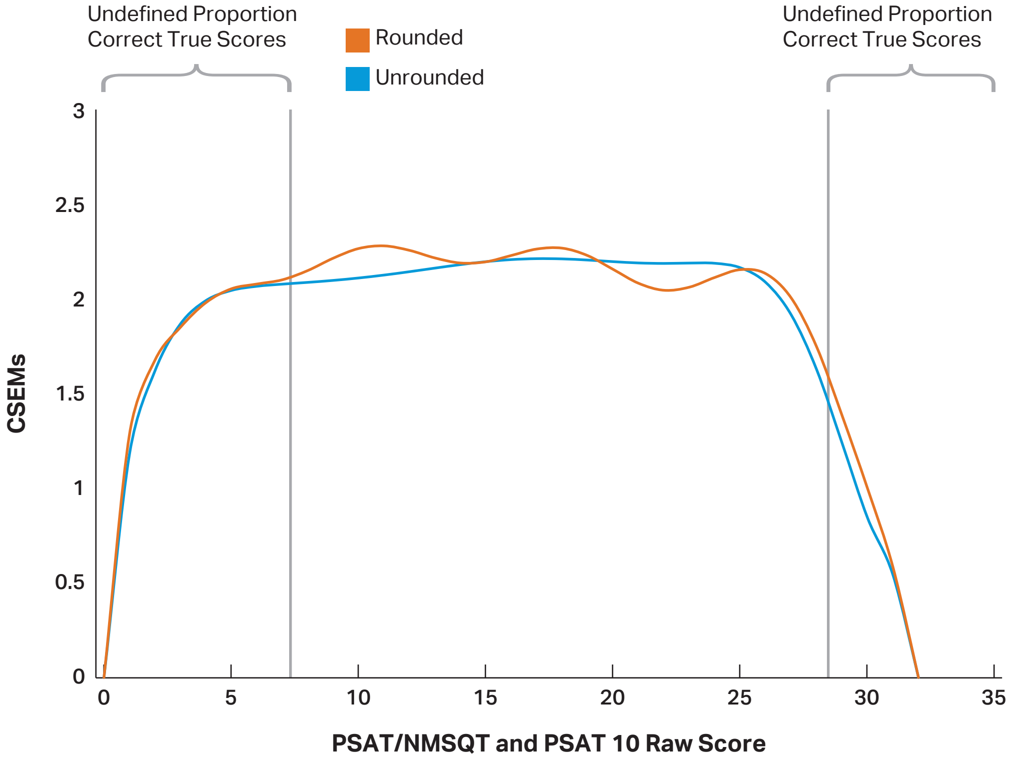


Figure 4.8: ***CSEMs*** **of the Adjusted, Rounded and Unrounded Scale Scores for PSAT 8/9 Reading (42 Items)**

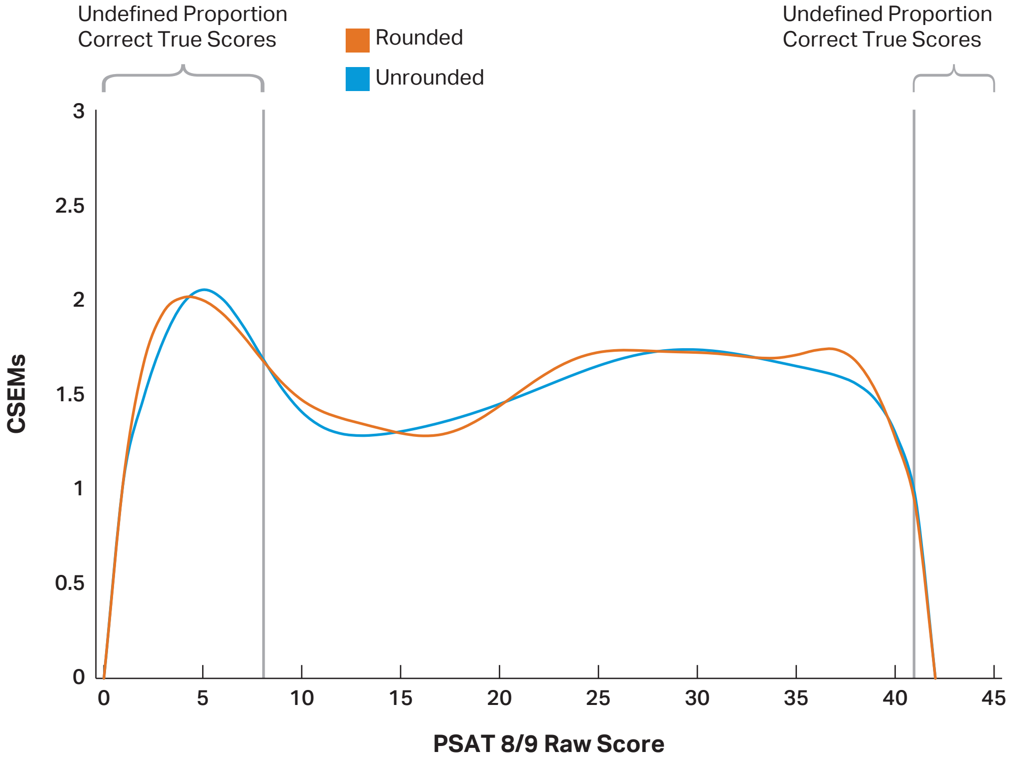


Figure 4.9: ***CSEMs*** **of the Adjusted, Rounded and Unrounded Scale Scores for PSAT 8/9 Math (38 Items)**

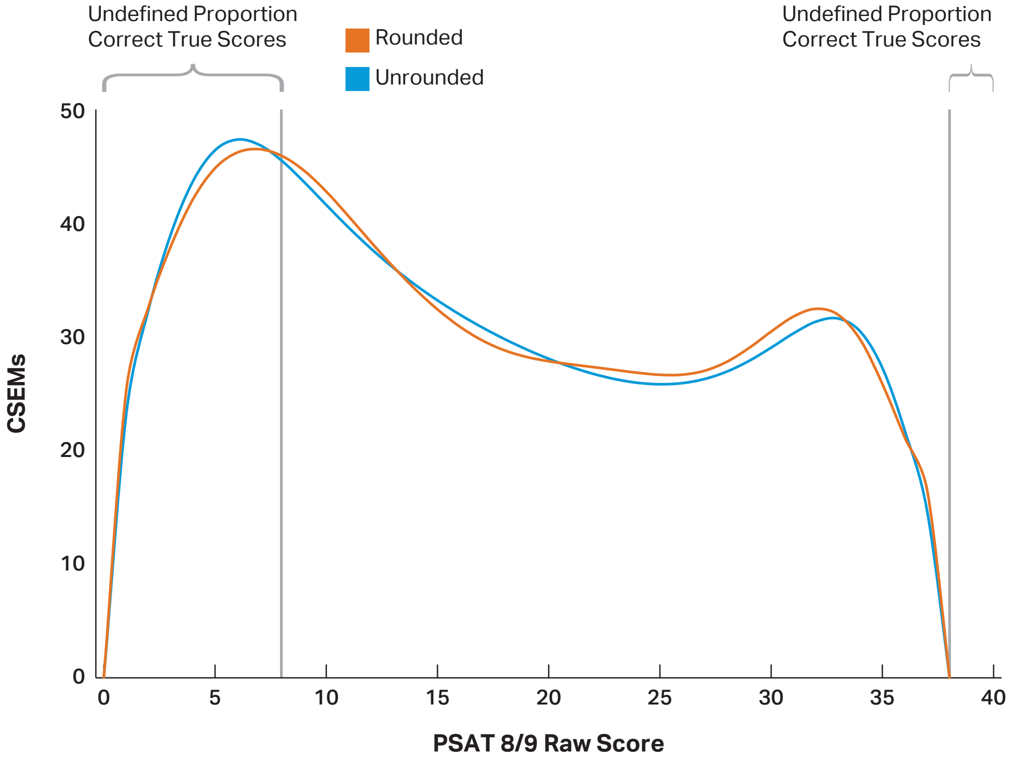


Figure 4.10: ***CSEMs*** **of the Adjusted, Rounded and Unrounded Scale Scores for PSAT 8/9 Writing and Language (40 Items)**

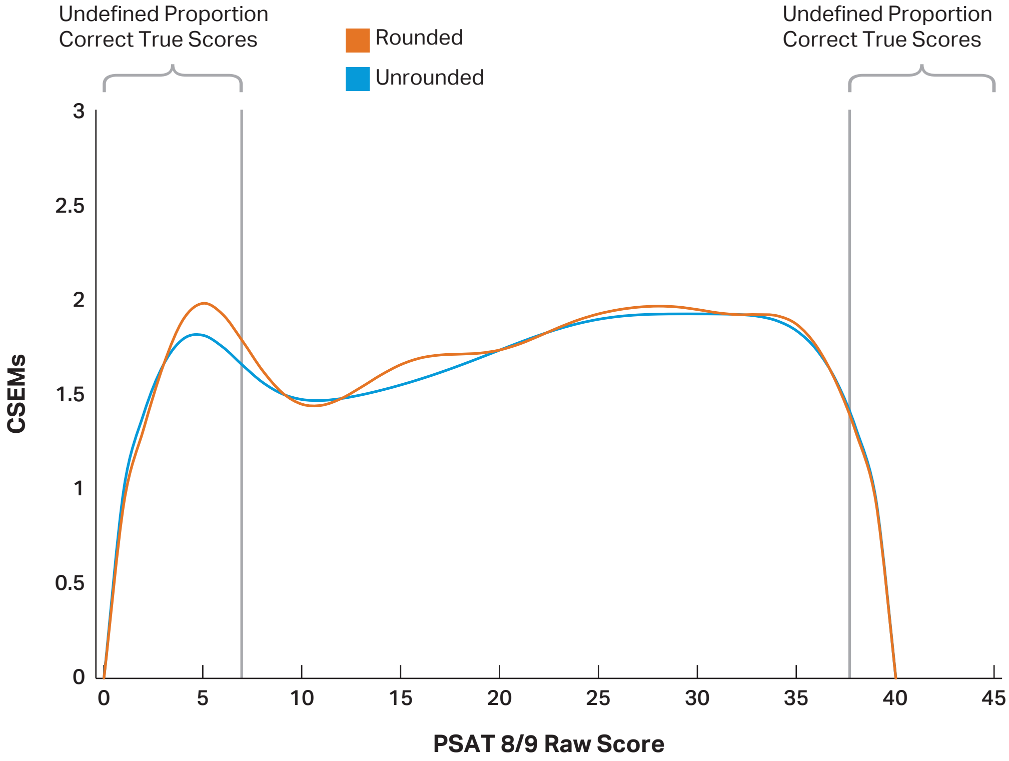


Figure 4.11: *CSEM*s of the Adjusted, Rounded and Unrounded Scale Scores for PSAT 8/9 Analysis in Science (29 Items)

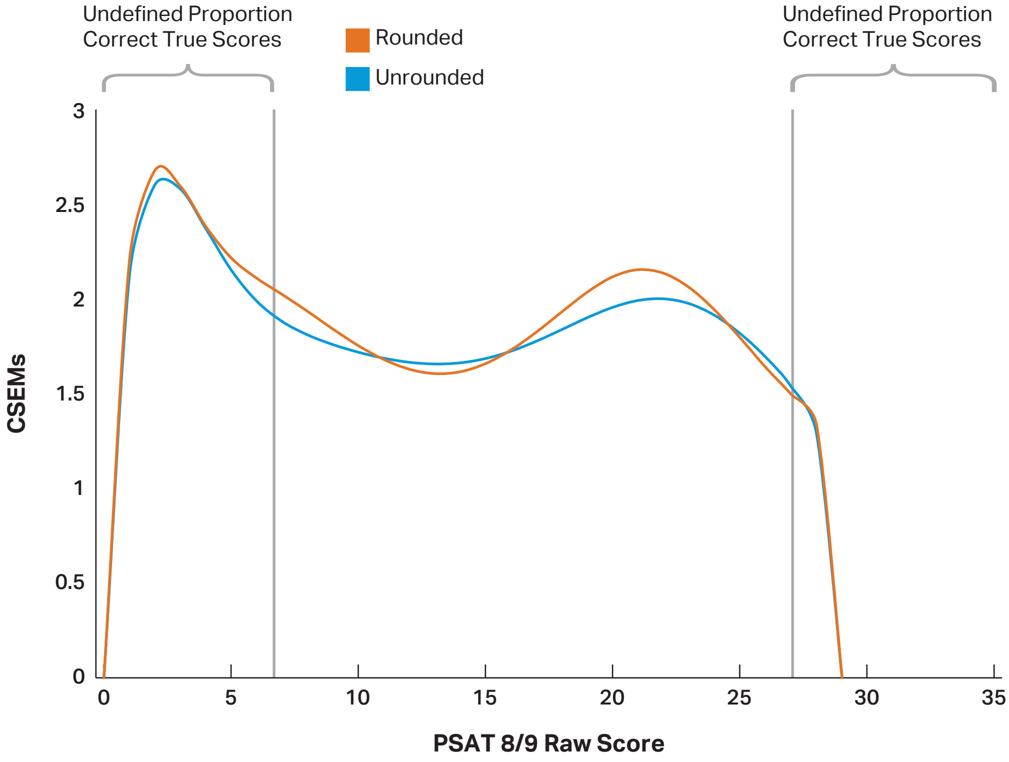
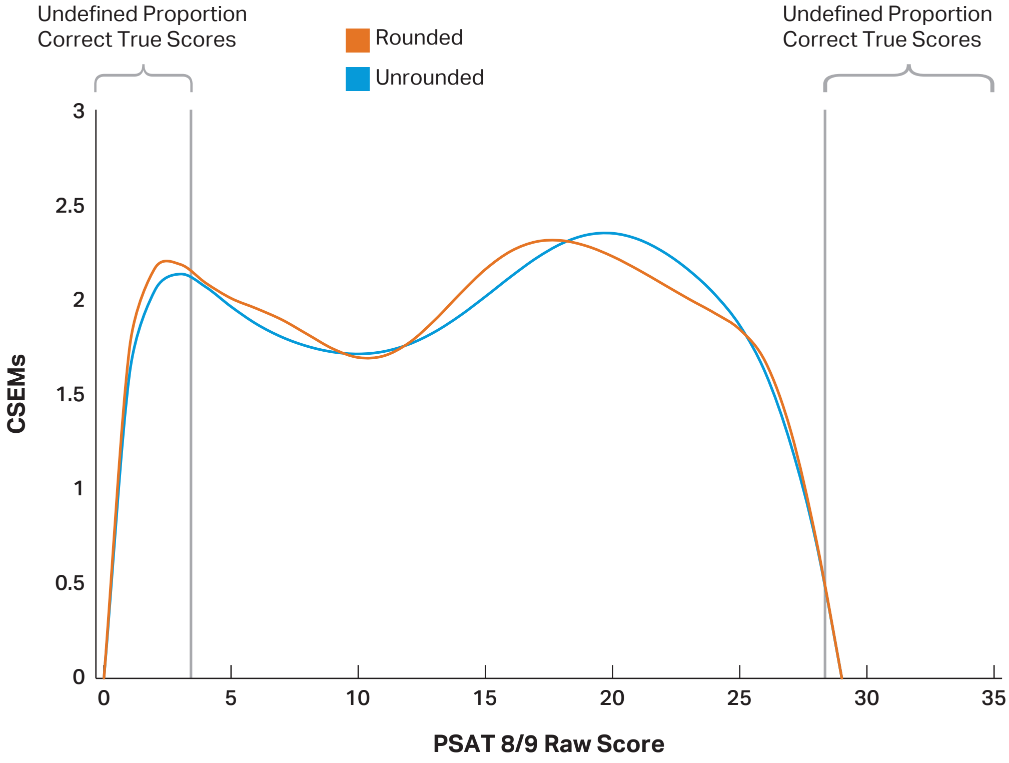


Figure 4.12: *CSEM*s of the Adjusted, Rounded and Unrounded Scale Scores for PSAT 8/9 Analysis in History/Social Studies (29 Items)



Discussion

The scaling decisions described in this chapter were made by utilizing the data in ways that would satisfy several criteria about approximating motivated test performance from the voluntary participants, approximating the nationally representative populations of interest, and producing scale scores with desired characteristics. These decisions were often made in connection with other decisions. For example, school recruitment, motivation screening, and sample weighting were implemented to approximate motivated, nationally representative groups with scale score distributions that lined up in expected ways. That is, the scale score distribution of the NRSAT student group should be higher than those of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 student groups. Scale score conversions for PSAT/NMSQT and PSAT 10 that reached the maximum possible scale scores for all test and cross-test scores except Writing and Language were established to reflect assumptions that the PSAT/NMSQT and PSAT 10 forms were more difficult than intended and likely to be more difficult than future forms. Smoothing and averaging methods were used with the equipercentile scaling to address statistical instability from scaling test subsamples that were smaller than anticipated. Results from all of these decisions had to work effectively with the previously established SAT scales (see Chapter 3). Different vertical scaling results would have likely been obtained if different data were obtained and/or different scaling decisions were made at any point in the process.

To the extent that the vertical scaling process described in this chapter was successful, these vertical scales support a vertically aligned longitudinal assessment system for evaluations of student growth across five domains and five grade levels. This means that scale scores obtained on the PSAT/NMSQT and PSAT 10 and PSAT 8/9 should be similar (ideally equivalent) to those that students might receive if they took the SAT rather than the PSAT/NMSQT and PSAT 10 or PSAT 8/9 on their testing date. Factors that affect how accurately the PSAT/NMSQT and PSAT 10 or PSAT 8/9 scale scores approximate SAT performance are differences in difficulty and content coverage of the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 tests; differences in the student groups taking each test; and differences in measurement precision (i.e., the PSAT/NMSQT and PSAT 10 and PSAT 8/9 tests are shorter and less reliable than the SAT tests). Readers should note that the kind of growth the vertical scales are intended to support is not actual growth in the sense of one particular student's expected SAT score compared to their PSAT/NMSQT and PSAT 10 or PSAT 8/9 score, as a student who goes on to take the SAT on a future date after taking PSAT/NMSQT and PSAT 10 and/or PSAT 8/9 tests can have different performance expectations due to growth over multiple testing occasions. Growth projections for the SAT Suite will be provided based on future studies.

BIBLIOGRAPHY/REFERENCES

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking. Methods and practices* (3rd ed.). New York, NY: Springer.

CHAPTER 5

Subscore Scaling—Characteristics of Subscore Scaling

YoungKoung Kim and Tim Moses

Subscores are intended for use by high schools to provide added insight about student achievement and to inform the assessment and improvement of students' college and career readiness and success. There are seven subscores that are intended to support the key features of the redesigned SAT Suite of Assessments (College Board, 2017):

- **Words in Context (WIC):** 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 honor the best work of the classroom. The skills tested are broadly useful in numerous subjects and careers. This subscore is composed of items from the Reading Test and the Writing and Language Test.
- **Command of Evidence (COE):** Test takers analyze material from a variety of content areas (literature and literary nonfiction, science, history, and social studies) and on career-related topics. Test takers use textual evidence to support their answers and apply an understanding of how authors make use of evidence. This subscore is composed of items from the Reading Test and the Writing and Language Test.
- **Expression of Ideas (EOI):** Addresses topic development, organization, and effective language use in Writing and Language.
- **Standard English Conventions (SEC):** Addresses sentence structure, usage, and punctuation in Writing and Language.
- **Math That Matters Most:** In keeping with the redesign's philosophy of a deeper focus on fewer topics, the Math Test contains subscores focused on three areas that reflect what research shows is essential for college readiness (College Board, 2016): Heart of Algebra (HOA), Problem Solving and Data Analysis (PSD), and Passport to Advanced Math (PAM). In addition to these three areas, the Math Test includes additional items focused on topics not included in these three areas.

The SAT and PSAT/NMSQT and PSAT 10 contain all seven subscores. The PSAT 8/9 includes the same subscores as the SAT and PSAT/NMSQT and PSAT 10 except for PAM. Subscores on the SAT, PSAT/NMSQT and PSAT 10, and the PSAT 8/9 were constructed independently of each other. The subscore scaling process for each assessment was consistent with the process for establishing the SAT test and cross-test scale scores, which was discussed in Chapter 3. This chapter describes the goals for establishing the subscore scales, the scaling process, and the results for the SAT, the PSAT/NMSQT and PSAT 10, and the PSAT 8/9.

Goals for the Subscore Scales

Scales were developed for the subscores of the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 using data from the 2014 scaling study. The scaling goals for the subscore scales can be summarized as follows:

SAT

- Ranges of 1–15
- Means of 8 for a college-bound group weighted to reflect old SAT cohorts
- Conditional Standard Errors of Measurement (*CSEM*s) that are approximately constant along the entire score range
- Standard deviations that are similar
- All correct, maximum possible raw scores convert to 15
- None correct, minimum possible raw scores convert to 1
- Minimized gaps and many-to-one conversions in the rounded scale scores

PSAT/NMSQT and PSAT 10 and PSAT 8/9

- Ranges of 1–15
- Means of 8 for a grade-specific nationally representative group
- *CSEM*s that are approximately constant along the entire score range
- Standard deviations that are similar
- All correct, maximum possible raw scores convert to 15
- None correct, minimum possible raw scores convert to 1
- Minimized gaps and many-to-one conversions in the rounded scale scores

Method

For the SAT, the subscore scaling used the same weighted sample as the one used to establish the SAT test and cross-test scale scores. After screening the examinees from the 2014 scaling study based on (1) completion rate, (2) survey questions on examinees' motivation, (3) grade level, and (4) educational plans beyond high school; the data were weighted to identify a college-bound group consisting of 11th and 12th graders. On the other hand, the subscore scalings of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 were based on the weighted samples for nationally representative groups. As discussed in Chapter 4, the scaling samples for the PSAT/NMSQT and PSAT 10 and PSAT 8/9 were composed of 9th and 10th graders, respectively.

The scales for the six subscores for the PSAT 8/9 and the seven subscores for the SAT and the PSAT/NMSQT and PSAT 10 were developed using methods that stabilize the *CSEM*s across the raw scores. Both arcsine transformation and cubic transformation methods were initially considered. As the final method, however, the arcsine transformation method was used to achieve constant *CSEM*s along the entire score range.

Subscore scaling used iterative scaling procedures similar to the ones employed for the SAT test scores and cross-test scores (see Chapter 3). Several *CSEM* levels were first examined and then 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. When the final methods were determined, the scaling methods that produced similar standard deviations across subscores were preferred. Linear interpolation adjustments (see Chapter 3, Equation 3.8) were applied to the highest and lowest scale scores to produce the 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 scaling described earlier in this chapter, the raw-to-rounded scale score conversions for the subscore scales were developed. The arcsine transformation method was used to set the scales for all subscores. The following desired *CSEM* values were selected for each subscore:

SAT

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

PSAT/NMSQT and PSAT 10

- Command of Evidence (COE): 1.3
- Words in Context (WIC): 1.4
- Standard English Conventions (SEC): 1.3
- Expression of Ideas (EOI): 1.1
- Heart of Algebra (HOA): 1.8
- Passport to Advanced Math (PAM): 1.6
- Problem Solving and Data Analysis (PSD): 1.4

PSAT 8/9

- Command of Evidence (COE): 1.4
- Words in Context (WIC): 1.5
- Standard English Conventions (SEC): 2.0
- Expression of Ideas (EOI): 1.2
- Heart of Algebra (HOA): 1.5
- Problem Solving and Data Analysis (PSD): 1.5

Table 5.1 shows the descriptive statistics for the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 subscore scales. Overall, the rounded scale score means for all subscores were very close to the target mean scores of 8. In addition, all scale scores appeared to have similar average *CSEM*s and similar standard deviations across all scale scores. Based on the

Table 5.1: Descriptive Statistics for the Subscore Scales

Assessment	Score	# of Items	Alpha Reliability	RMS (CSEM)	Scale Score Reliability from RMS (CSEM)	Mean	SD	Skew	Kurt	Min Possible (Observed)	Max Possible (Observed)
SAT	COE	18	0.76	1.3	0.75	8.0	2.6	0.36	−0.5	1 (1)	15 (15)
	WIC	18	0.74	1.6	0.75	8.2	3.2	−0.25	−0.7	1 (1)	15 (15)
	SEC	20	0.80	1.4	0.81	8.0	3.3	0.09	−0.6	1 (1)	15 (15)
	EOI	24	0.83	1.2	0.82	7.9	2.8	0.19	−0.6	1 (1)	15 (15)
	HOA	19	0.75	1.4	0.74	8.0	2.7	0.35	−0.2	1 (1)	15 (15)
	PAM	16	0.70	1.6	0.69	7.8	2.9	0.37	−0.1	1 (1)	15 (15)
	PSD	17	0.80	1.4	0.81	8.2	3.2	−0.13	−0.80	1 (1)	15 (15)
PSAT10	COE	18	0.71	1.3	0.70	8.0	2.4	0.16	−0.31	1 (1)	15 (15)
	WIC	18	0.76	1.3	0.76	8.0	2.7	0.28	−0.41	1 (1)	15 (15)
	SEC	20	0.74	1.3	0.73	8.0	2.5	0.24	−0.23	1 (2)	15 (15)
	EOI	24	0.79	1.1	0.78	8.0	2.3	0.40	0.16	1 (2)	15 (15)
	HOA	16	0.62	1.7	0.63	7.7	2.8	0.61	0.07	1 (1)	15 (15)
	PAM	14	0.61	1.7	0.54	8.0	2.6	−0.18	0.39	2 (2)	15 (15)
	PSD	16	0.65	1.4	0.65	7.9	2.4	0.18	0.18	1 (1)	15 (15)
PSAT8/9	COE	18	0.71	1.4	0.70	7.9	2.5	0.07	−0.3	1 (1)	15 (15)
	WIC	18	0.78	1.5	0.79	8.1	3.2	0.05	−0.6	1 (1)	15 (15)
	SEC	16	0.63	1.8	0.64	7.9	3.1	0.09	−0.3	1 (1)	15 (15)
	EOI	24	0.81	1.2	0.80	8.0	2.6	0.12	−0.5	1 (1)	15 (15)
	HOA	16	0.68	1.5	0.67	8.0	2.5	0.39	0.0	1 (1)	15 (15)
	PAM	—	—	—	—	—	—	—	—	—	—
	PSD	16	0.69	1.4	0.69	8.0	2.6	0.42	0.0	1 (1)	15 (15)

methods described in Chapter 3, the estimated scale score reliabilities for the seven SAT subscores ranged between 0.69 and 0.82. The estimated scale score reliabilities for the seven PSAT/NMSQT and PSAT 10 subscores ranged between 0.54 and 0.78. Lastly, the estimated scale score reliabilities for the six PSAT 8/9 subscores ranged between 0.64 and 0.80.

The descriptive tables show that the subscore criteria were less difficult to meet for the non-math subscores than for the math subscores. The COE, WIC, SEC, and EOI subscores contained more items than possible scale score points, had distributions that were usually not extremely skewed, and had subscore scales that met the subscore scaling criteria. The HOA, PAM, and PSD subscores of the Math Test had fewer items and, especially for the PSAT/NMSQT and PSAT 10, were obtained from a relatively difficult form so that the scale score mean of 8 resulted in gaps in conversion tables at the bottom range of the scales and more raw scores transformed into scale scores of 15 at the top range of the scales. Although different scaling methods were considered based on setting lower means for the HOA and PAM subscores of the PSAT/NMSQT and PSAT 10, these were ultimately not used because the means of 8 were regarded as a higher priority than minimizing gaps.

The plots of the adjusted unrounded and rounded scale score *CSEM*s for the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 subscores are presented in Figures 5.1–5.6. As shown in Figures 5.1, 5.3, and 5.5, the unrounded scale score *CSEM*s appeared to be constant across all subscores for the three assessments. On the other hand, the *CSEM*s of adjusted and rounded scale scores for some subscores were slightly inconsistent for certain ranges of scale scores mainly due to the adjustment for the highest and lowest scale scores, truncation and rounding of the unrounded scale scores, and the presence of a smaller number of items compared to test and cross-test scores (See Figures 5.2, 5.4, and 5.6).

Figure 5.1: *CSEM*s of the Adjusted, Unrounded Scale Scores for SAT Subscores

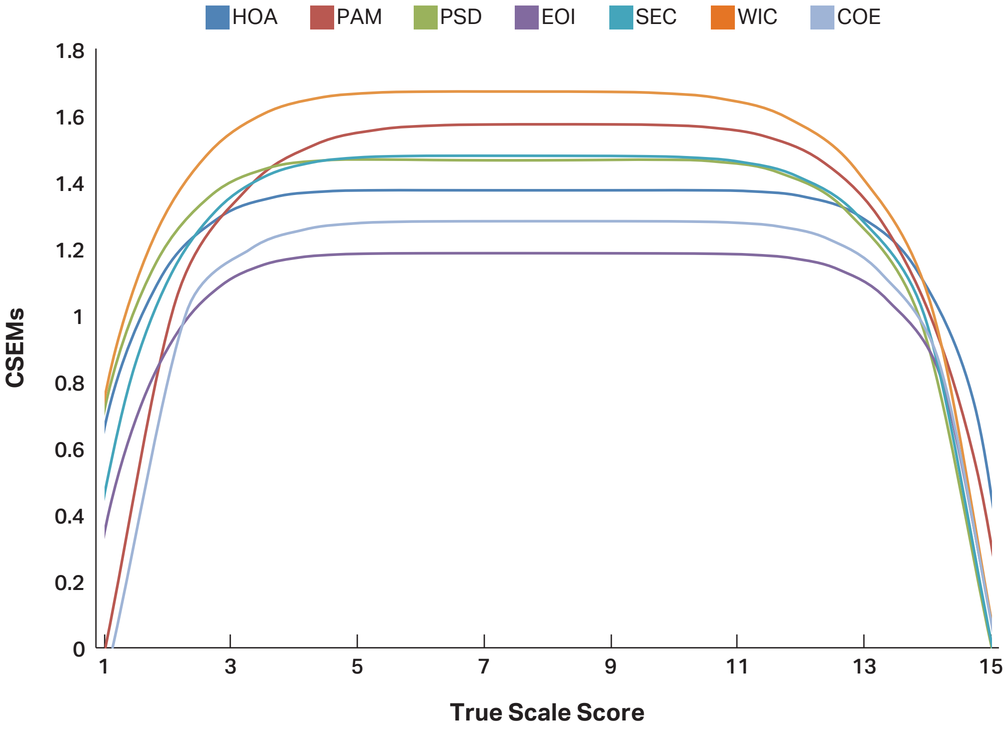


Figure 5.2: ***CSEM*s of the Adjusted, Rounded Scale Scores for SAT Subscores**

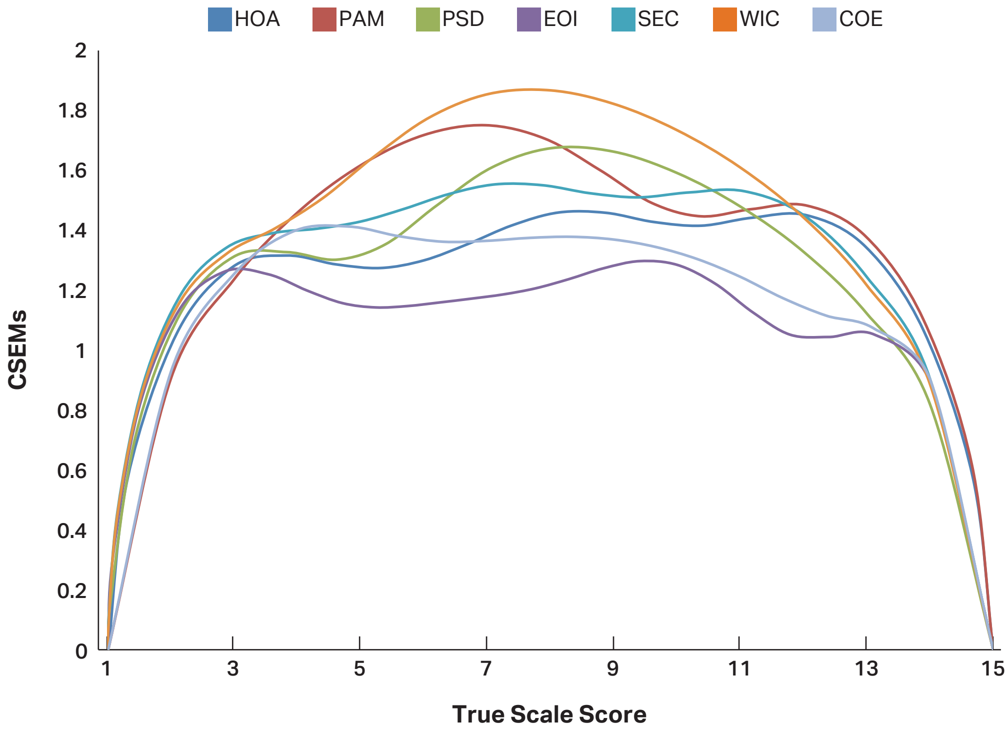


Figure 5.3: ***CSEM*s of the Adjusted, Unrounded Scale Scores for PSAT/NMSQT and PSAT 10 Subscores**

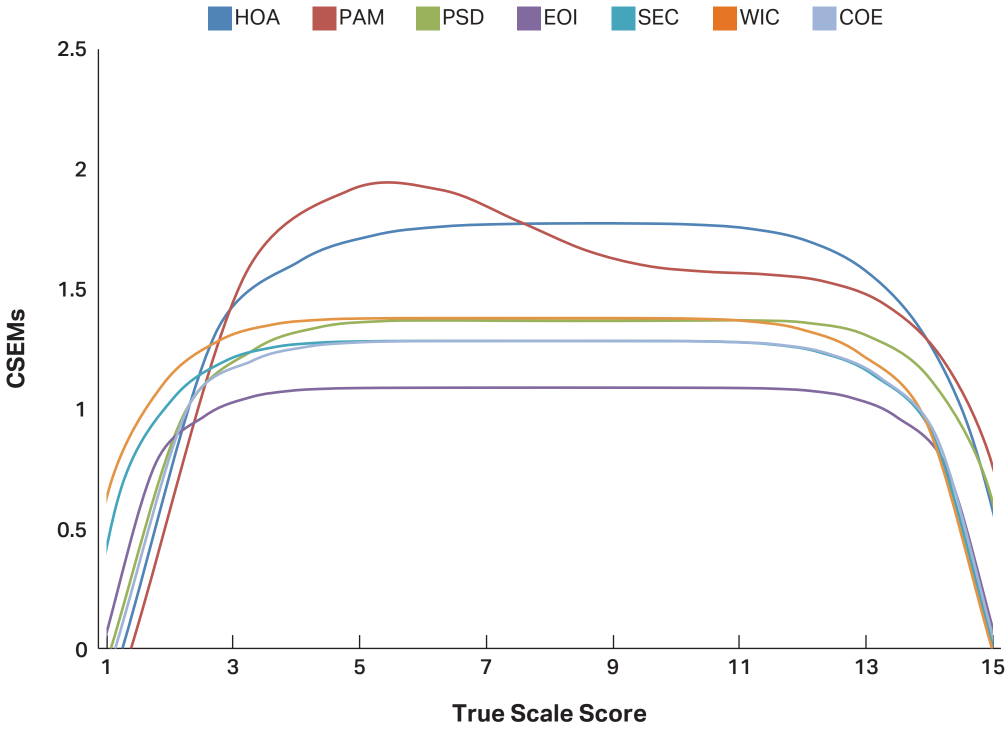


Figure 5.4: ***CSEMs*** **of the Adjusted, Rounded Scale Scores for PSAT/NMSQT and PSAT 10 Subscores**

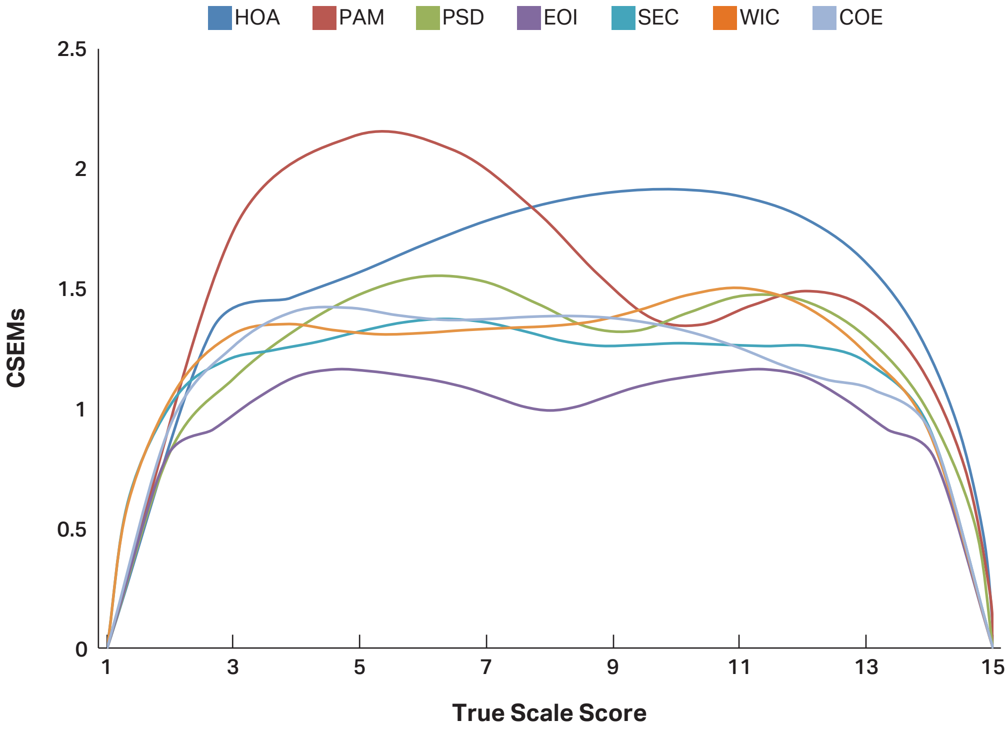


Figure 5.5: ***CSEMs*** **of the Adjusted, Unrounded Scale Scores for PSAT 8/9 Subscores**

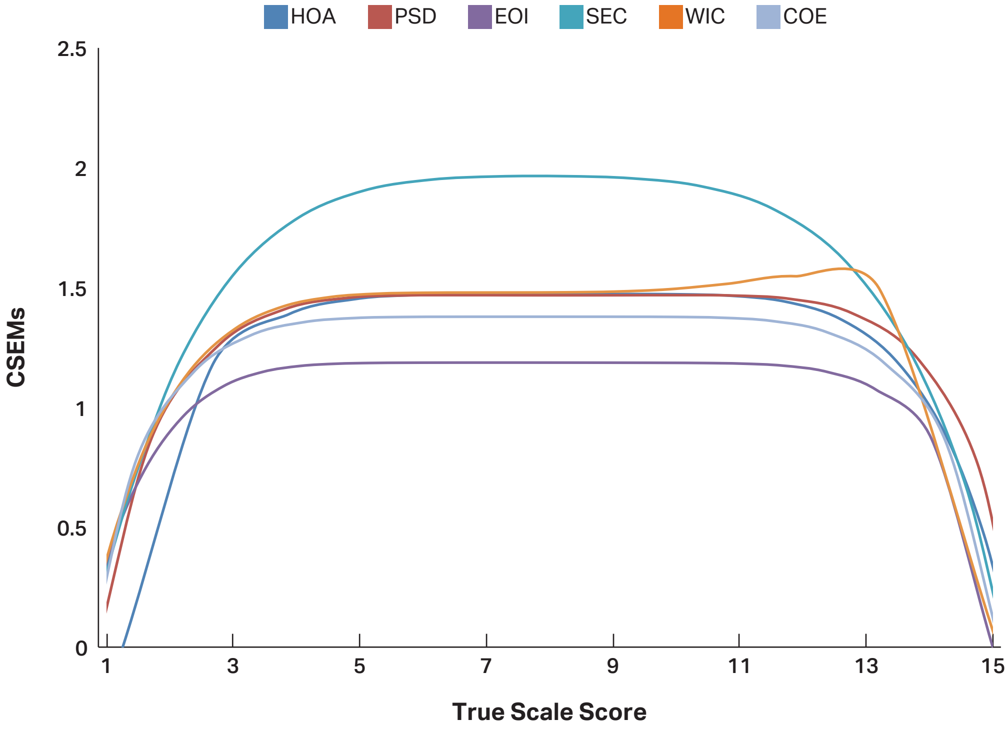
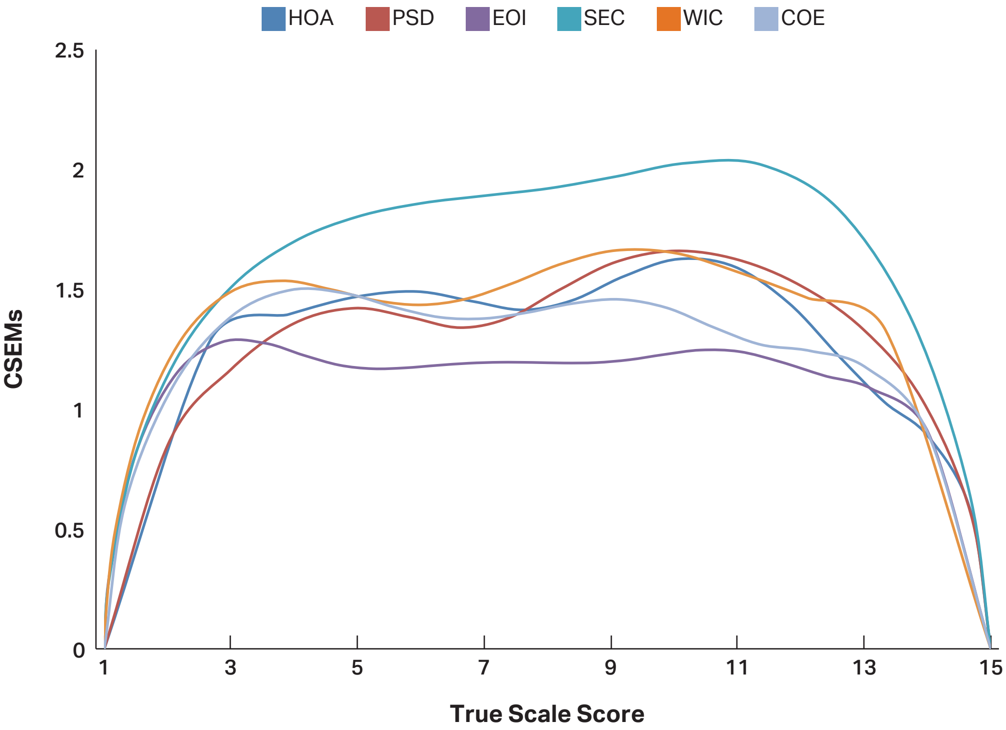


Figure 5.6: ***CSEM*****s of the Adjusted, Rounded Scale Scores for PSAT 8/9 Subscores**



Evaluations of the Subscore Scales

The results of the subscore scalings were generally reflective of the scaling goals, in that subscore scales ranging from 1–15 with means of approximately 8 and with reasonably stabilized *CSEM*s were obtained for most of the 20 (7 SAT + 7 PSAT/NMSQT and PSAT 10 + 6 PSAT 8/9) subscores. For some subscores, the scaling goals were more difficult to meet. In particular, the PAM subscore for the PSAT/NMSQT and PSAT 10 was based on 14 items and was relatively difficult, and setting the mean scale score at 8 resulted in a stretching of the lowest scale scores, more gaps, and relatively large *CSEM*s for these low scores (Figure 5.3 and 5.4). Possibly, more stable *CSEM*s might have been achieved for PAM in the PSAT/NMSQT and PSAT 10 if the scale score mean was set lower than 8, which would have been more closely reflective of the difficulty of this subscore and might have resulted in less stretching of the lowest scores and less inflation of the *CSEM*s. These issues were considered in reviews of all the subscore scaling results, and final decisions were made to achieve consistency with respect to targeted scale score means.

Because the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 subscores were produced from the same datasets as the SAT scales (the sample that approximated the SAT cohort; Chapter 3) and the vertical scales (samples that approximated nationally representative students; Chapter 4), several of the cautions described for the SAT scale and vertical scales also apply to the subscore scaling results. That is, the uncertainty in the representativeness of the average performance in the scaling study data with respect to operational SAT test taker performance implied that the means of the SAT subscores may not be exactly 8 when taken by operational SAT test takers. Likewise, the weighted and cleaned datasets used for the vertical scalings of the PSAT/NMSQT and PSAT 10 and PSAT 8/9 implied that the means of these subscores may not be 8 when taken by operational PSAT/NMSQT and PSAT 10 and PSAT 8/9 test takers. At the time of this writing, the scales have been monitored with respect to initial administrations of the redesigned SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9, and results have been studied with respect to equating results. Further monitoring of these results will be useful for informing times in which the subscore scales may need alterations to more clearly meet the goals of the subscores to provide useful added insight into student achievement.

BIBLIOGRAPHY/REFERENCE

College Board. (2017). *SAT Suite of Assessments technical manual: Characteristics of the SAT*. New York, NY: The College Board.

CHAPTER 6

Discussion

Tim Moses

A major question about the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 scaling results described in this monograph was, “Were the scaling goals met?” From the perspective of the 2014 scaling study data, the answer to this question is “yes.” That is, the schools and students involved in the study were appropriately targeted and nationally representative, were screened for motivation levels and weighted for demographic representativeness that appeared to reflect the target populations for the scalings, and the resulting data were used to set scales that met pre-established scaling criteria. The resulting scales had the desired minimum scores, maximum scores, similar distributions, relatively stabilized *CSEM*s, and, for the data worked with, the targeted means and standard deviations. The scale scores appear to have similar features (i.e., ranges, means, distributions) at the section level, the test and cross-test levels, and the subscore level, which allows for informative evaluations of performance across content domains. Additional analyses of equatings of other SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 test forms to the scales suggested that the desired scaling properties would be reasonably maintained in subsequently developed test forms for which SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 scores would be reported (College Board, 2017). Of course, scale score characteristics will be monitored over time based on the performance of operationally tested examinees (Standard 5.6, AERA/APA/NCME, 2014).

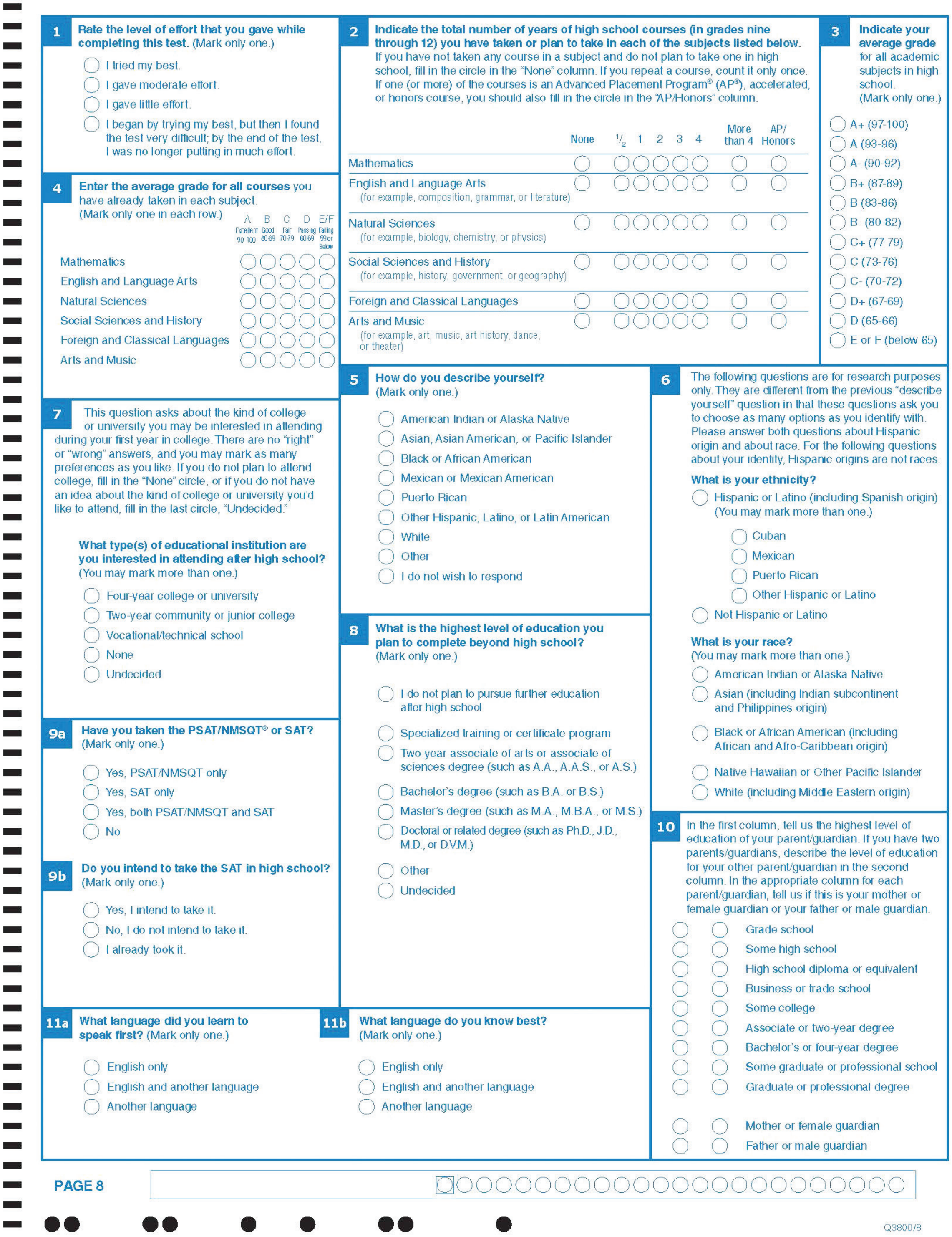
Ultimately, the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9 scales are intended to be useful for the multiple purposes sought in the redesigned SAT Suite of Assessments (see Chapter 1). The scales are intended to be accurate indicators of college and career readiness, with the capacity to have distinguishable scores and ranges from which it is possible to predict success in first-year college courses (College Board, 2017). The scales should have usefulness for reporting on the performances for multiple populations, such as those of college-bound graduating seniors traditionally tracked by the College Board, athletes seeking college admission and scholarships (NCAA), 11th-grade PSAT/NMSQT test takers seeking National Merit recognition, and nationally representative and state-specific students at specific grade levels for the SAT, PSAT/NMSQT and PSAT 10, and PSAT 8/9. Finally, the vertical scales for the PSAT/NMSQT and PSAT 10 and the PSAT 8/9 are intended to be used as meaningful indications of SAT performance, in that they are a starting point from which measures of growth might be studied, developed, and reported at various high school grades for the score levels and content domains measured by the SAT Suite.

BIBLIOGRAPHY/REFERENCES

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

College Board. (2017). *SAT Suite of Assessments technical manual: Characteristics of the SAT.* New York, NY: The College Board.

APPENDIX



1 **Rate the level of effort that you gave while completing this test.** (Mark only one.)

○ I tried my best.
○ I gave moderate effort.
○ I gave little effort.
○ I began by trying my best, but then I found the test very difficult; by the end of the test, I was no longer putting in much effort.

2 **Indicate the total number of years of high school courses (in grades nine through 12) you have taken or plan to take in each of the subjects listed below.** If you have not taken any course in a subject and do not plan to take one in high school, fill in the circle in the "None" column. If you repeat a course, count it only once. If one (or more) of the courses is an Advanced Placement Program® (AP®), accelerated, or honors course, you should also fill in the circle in the "AP/Honors" column.

	None	1/2	1	2	3	4	More than 4	AP/ Honors
Mathematics	○	○	○	○	○	○	○	○
English and Language Arts (for example, composition, grammar, or literature)	○	○	○	○	○	○	○	○
Natural Sciences (for example, biology, chemistry, or physics)	○	○	○	○	○	○	○	○
Social Sciences and History (for example, history, government, or geography)	○	○	○	○	○	○	○	○
Foreign and Classical Languages	○	○	○	○	○	○	○	○
Arts and Music (for example, art, music, art history, dance, or theater)	○	○	○	○	○	○	○	○

3 **Indicate your average grade** for all academic subjects in high school. (Mark only one.)

○ A+ (97-100)
○ A (93-96)
○ A- (90-92)
○ B+ (87-89)
○ B (83-86)
○ B- (80-82)
○ C+ (77-79)
○ C (73-76)
○ C- (70-72)
○ D+ (67-69)
○ D (65-66)
○ E or F (below 65)

4 **Enter the average grade for all courses** you have already taken in each subject. (Mark only one in each row.)

	A Excellent 90-100	B Good 80-89	C Fair 70-79	D Passing 60-69	E/F Failing 59 or Below
Mathematics	○	○	○	○	○
English and Language Arts	○	○	○	○	○
Natural Sciences	○	○	○	○	○
Social Sciences and History	○	○	○	○	○
Foreign and Classical Languages	○	○	○	○	○
Arts and Music	○	○	○	○	○

5 **How do you describe yourself?** (Mark only one.)

○ American Indian or Alaska Native
○ Asian, Asian American, or Pacific Islander
○ Black or African American
○ Mexican or Mexican American
○ Puerto Rican
○ Other Hispanic, Latino, or Latin American
○ White
○ Other
○ I do not wish to respond

6 The following questions are for research purposes only. They are different from the previous "describe yourself" question in that these questions ask you to choose as many options as you identify with. Please answer both questions about Hispanic origin and about race. For the following questions about your identity, Hispanic origins are not races.

What is your ethnicity?

○ Hispanic or Latino (including Spanish origin) (You may mark more than one.)
 ○ Cuban
 ○ Mexican
 ○ Puerto Rican
 ○ Other Hispanic or Latino
○ Not Hispanic or Latino

What is your race? (You may mark more than one.)

○ American Indian or Alaska Native
○ Asian (including Indian subcontinent and Philippines origin)
○ Black or African American (including African and Afro-Caribbean origin)
○ Native Hawaiian or Other Pacific Islander
○ White (including Middle Eastern origin)

7 This question asks about the kind of college or university you may be interested in attending during your first year in college. There are no "right" or "wrong" answers, and you may mark as many preferences as you like. If you do not plan to attend college, fill in the "None" circle, or if you do not have an idea about the kind of college or university you'd like to attend, fill in the last circle, "Undecided."

What type(s) of educational institution are you interested in attending after high school? (You may mark more than one.)

○ Four-year college or university
○ Two-year community or junior college
○ Vocational/technical school
○ None
○ Undecided

8 **What is the highest level of education you plan to complete beyond high school?** (Mark only one.)

○ I do not plan to pursue further education after high school
○ Specialized training or certificate program
○ Two-year associate of arts or associate of sciences degree (such as A.A., A.A.S., or A.S.)
○ Bachelor's degree (such as B.A. or B.S.)
○ Master's degree (such as M.A., M.B.A., or M.S.)
○ Doctoral or related degree (such as Ph.D., J.D., M.D., or D.V.M.)
○ Other
○ Undecided

9a **Have you taken the PSAT/NMSQT® or SAT?** (Mark only one.)

○ Yes, PSAT/NMSQT only
○ Yes, SAT only
○ Yes, both PSAT/NMSQT and SAT
○ No

9b **Do you intend to take the SAT in high school?** (Mark only one.)

○ Yes, I intend to take it.
○ No, I do not intend to take it.
○ I already took it.

10 In the first column, tell us the highest level of education of your parent/guardian. If you have two parents/guardians, describe the level of education for your other parent/guardian in the second column. In the appropriate column for each parent/guardian, tell us if this is your mother or female guardian or your father or male guardian.

○	○	Grade school
○	○	Some high school
○	○	High school diploma or equivalent
○	○	Business or trade school
○	○	Some college
○	○	Associate or two-year degree
○	○	Bachelor's or four-year degree
○	○	Some graduate or professional school
○	○	Graduate or professional degree
○	○	Mother or female guardian
○	○	Father or male guardian

11a **What language did you learn to speak first?** (Mark only one.)

○ English only
○ English and another language
○ Another language

11b **What language do you know best?** (Mark only one.)

○ English only
○ English and another language
○ Another language

PAGE 8

Q3800/8