RESEARCH REPORT 2012-3

CollegeBoard

A Standard-Setting Study to Establish College Success Criteria to Inform the SAT® College and Career Readiness Benchmark

By Jennifer L. Kobrin, Brian F. Patterson, Andrew Wiley, and Krista D. Mattern

Jennifer L. Kobrin is a research scientist at the College Board.

Brian F. Patterson is an assistant research scientist at the College Board.

Andrew Wiley is the executive director of assessment design and test development at the College Board.

Krista D. Mattern is an associate research scientist at the College Board.

Mission Statement

The College Board's mission is to connect students to college success and opportunity. We are a not-for-profit membership organization committed to excellence and equity in education.

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 www.collegeboard.org.

VALIDITY

For more information on College Board research and data, visit www.collegeboard.org/research.

Contents

Tables

Figures

Executive Summary

In 2011, the College Board released its SAT® college and career readiness benchmark[1], which represents the level of academic preparedness associated with a high likelihood of college success and completion. The goal of this study, which was conducted in 2008, was to establish college success criteria to inform the development of the benchmark. The College Board convened a panel comprised of experts in educational policy and higher education to review data showing the relationship between SAT scores and college performance. Panelists were asked to provide two sets of ratings on what first-year college GPA (FYGPA) should be used to define the criterion for success in the first year of college; and two sets of ratings to define the probability level for a successful student attaining that FYGPA (probability of mastery). The mean FYGPA rating from the second round was 2.62 (with a median of 2.67), and the mean and median rating for probability of mastery was 70%. The SAT score associated with the panelists' final ratings was approximately 1580.

1. The SAT college and career readiness benchmark of 1550, described in Wyatt, Kobrin, Wiley, Camara, and Proestler (2011), was calculated as the SAT score associated with a 65 percent probability of earning a first-year GPA of 2.67 (B-) or higher. The probability level of 65 percent was ultimately chosen because it was within the range of panelists' recommendations and has been used by NAEP and in other educational settings.

Introduction

The College Board's mission is to help connect students to college. One important step in fulfilling this mission is gauging if students have the academic skills needed to succeed in college. Key aspects of college readiness are the knowledge and skills students learn in school, along with the academic skills (such as reasoning, problem solving, and writing abilities) as demonstrated by successful performance on the SAT (Wiley, Wyatt, & Camara, 2010). The SAT is a standardized assessment of the critical reading, mathematics, and writing skills students have developed during their academic careers. Students' scores on each of the three sections of the SAT range from 200 to 800, for a combined total score ranging from 600 to 2400. The average score on each section is approximately 500. Each year, more than two million students take the SAT, and nearly every four-year college in the United States uses the test as a common and objective scale for evaluating students' college readiness.

While school districts and state departments of education have access to the average SAT scores of their students and could examine trends over the years, prior to 2011 there was no point of reference to help these educators and policymakers determine what proportion of their students were actually prepared to succeed in college. The SAT college and career readiness benchmark was developed to help secondary school administrators, educators, and policymakers evaluate the effectiveness of academic programs in order to better prepare students for success in college and beyond (College Board, 2011). The College Board continues to advocate that SAT scores be used in combination with other indicators, such as high school GPA, when making high-stakes decisions regarding an individual's college readiness.

The Importance of Benchmarks

Today, a large percentage of high school students have aspirations to attend college, but only approximately half of the students who enroll in college are actually prepared for college-level academic work (Kirst & Venezia, 2006). This discrepancy highlights the critical need to inform educators and administrators of the proportion of their students with the academic skills to succeed in college, and to help evaluate the effectiveness of their academic programs in preparing their students for postsecondary opportunities. Top policy groups, such as the National Governors Association, the Council of Chief State School Officers, and the standards-advocacy group Achieve, are pushing states toward benchmarking as a way to better prepare students for a competitive global economy (McNeil, 2008). The use of benchmarks is also being recommended at the local school level to gauge the college readiness of individual schools' students. In the report *Getting There — and Beyond: Building a Culture of College-going in High Schools*, the University of Southern California Center for Higher Education Policy Analysis (Corwin & Tierney, 2007), it is suggested that schools create benchmarks to strengthen the college culture and expectations for students. For example, the report recommends the creation of benchmarks for high school juniors, and a system that would verify that juniors are on track to complete college eligibility requirements during their senior year.

States are increasingly attempting to incorporate the concept and rigor associated with college readiness in both their content-based state standards and their student-level descriptions.

States are increasingly attempting to incorporate the concept and rigor associated with college readiness in both their content-based state standards and their student-level descriptions. The Common Core standards initiative and proposed assessments represent a major effort by states to establish consistent content and performance standards related to college readiness. Several states have also incorporated empirically based benchmarks in setting cut scores on state tests to ensure college readiness.

Current Conceptions of College Readiness and College Success

There are a range of opinions about what it means to be ready for college and what it means to be successful in college. Achieve, Inc., in partnership with the Education Trust and the Thomas B. Fordham Foundation, took a content-driven approach in their American Diploma Project (ADP) (Achieve, Inc., The Education Trust, & Thomas B. Fordham Foundation, 2004). The ADP focused on codifying the English and mathematics skills that high school graduates need in order to be successful in college and the workplace. The ADP and partner organization staff worked with faculty members of two- and four-year institutions to define the content and skills necessary for success in freshman credit-bearing courses. They also used data from the Bureau of Labor Statistics and the U.S. Department of Education's National Education Longitudinal Study (NELS) to identify "good" jobs — those that pay enough to support a family above the poverty level, provide benefits, and offer advancement. The high school courses taken by individuals in these "good" jobs and the grades achieved in these courses were recorded. The postsecondary and workplace expectations were combined into a set of ADP college and workplace readiness benchmarks. In math, the benchmarks contain content typically taught in Algebra I, Algebra II, geometry, and data analysis and statistics. In English, the benchmarks require strong oral and written communication skills that are essential in college classrooms and high-performance workplaces. These benchmarks also describe analytical and research skills currently associated only with advanced and honors courses.

There are a range of opinions about what it means to be ready for college and what it means to be successful in college.

David Conley, with support from the Gates Foundation, provided a comprehensive definition of college readiness that includes key cognitive strategies, key content, academic behaviors, and contextual skills and awareness (Conley, 2007). Conley presented a representative list of the knowledge, skills, and attributes a student should possess to be ready to succeed in entry-level college courses across a range of subjects and disciplines, and provided example performances of students who have acquired the necessary competence in these domains.

The Southern Regional Education Board (SREB) (2002) embarked on the College Readiness Policy Connections initiative designed to highlight student preparation for college and careers, and to help states identify policy gaps and weaknesses that may hinder their students from reaching their college potential. The SREB and its three partner states (Georgia, Texas, and West Virginia) identified 24 student needs associated with college readiness. These needs fell into the following areas: curriculum and standards, assessment and accountability, educational support systems, qualified professional staff, community and parental partnerships, and facilities, equipment and instructional materials.

Other researchers and organizations have taken a more empirical approach to establishing criteria of college readiness and college success. In addition to its content-oriented standards for college success described above, the SREB (Lord, 2003) also defined college readiness indices using college admission tests. Following the National Assessment of Educational Progress (NAEP) performance levels, SREB defined four categories of college readiness: 1) Basic, 2) Admissible, 3) Standard, and 4) Proficient. Of the Basic category, the report stated that students in this category are "generally sufficient for admission to degree programs at non-selective institutions, but students with these scores are generally required to take remedial courses." (page 15). Of the Proficient category, the report stated that these students are prepared for admission to selective programs like engineering, or for admission to selective or competitive institutions. In 2002, the percentage of students in SREB states meeting the benchmarks was 80%–85% for the Basic category, 65%–71% for the Admissible category, 46%–57% for the Standard category, and 16%–26% for the Proficient category. The report advised that when evaluating the percentage of students meeting the benchmarks, it is important to consider the proportion of high school seniors taking the tests in each state, since not all students take college admission tests.

Greene and Winters (2005) calculated a measure of public high school college readiness designed to reproduce the minimum standards of the least selective four-year colleges. The standards included earning a regular high school diploma, completing a minimum set of course requirements, and being able to read at a basic level (scoring at or above the basic level on the National Assessment of Educational Progress reading assessment). According to their measure of college readiness, Greene and Winters estimated that only 34% of 2002 high school graduates in the nation had the skills and qualifications necessary to attend college. The New England Board of Higher Education (2006) used the Greene and Winters measure to determine the college readiness of students in New England states, and noted a significant gap in college readiness for underrepresented minority students.

The National Center for Education Statistics' (NCES) measure of college readiness was based on a student's cumulative grades in high school academic course work, senior class rank, the National Education Longitudinal Study (NELS) 1992 test scores, and college entrance examination scores (Berkner & Chavez, 1997). Each student was rated on a five-point scale, ranging from "marginally or not qualified" to "very highly qualified," based on his/her highest-rated criterion. In addition, students were moved up one category if they took rigorous academic course work (at least four years of English; three years each of a natural science, social science, and math; and two years of a foreign language) and demoted one category if they did not take such course work. According to this college qualification index, among all 1992 high school graduates nearly two-thirds (65%) appeared to have been at least minimally qualified for admission to a four-year college or university. Among those seniors classified as marginally or not qualified for regular four-year college admission, half entered postsecondary education, but only 15% enrolled in a four-year college or university. Among those seniors who were minimally qualified, three-quarters enrolled in some postsecondary education, and 35% attended a four-year institution. Fifty-six percent of the somewhat qualified, 73% of the highly qualified, and 87% of the very highly qualified high school graduates enrolled in four-year institutions.

Twing, Miller, and Meyers (2008) conducted a standard-setting study to determine a performance standard on the Texas high school assessment. This performance standard was intended to identify students likely to be ready for success with college-level work, and was intended for use as a cut score required for students in Texas to receive college instruction. In setting this performance standard, the criteria used to indicate whether a student was successful in college included: returned to college for a second semester, earned a college GPA of no less than 2.0, and took no remedial course work in the first semester of college.

Montgomery County in Maryland developed a measure of college readiness based on seven key indicators, which include advanced reading achievement in grades K–8; completion of grade 6 math by grade 5; completion of Algebra I by grade 8, and Algebra II by grade 11; minimum scores on at least one AP or IB exam, and an SAT combined score of 1650 or higher, or an ACT composite score of 24 or higher (Von Secker, 2009).

In their report, *Crisis at the Core* (2004), ACT indicated that most of America's high school students are not ready for college-level course work. Using the criteria of a 75% chance of earning a grade of C or better and a 50% chance of earning a B or better in first-year college English composition, algebra, and biology courses, only 68% of ACT-tested high school graduates met the benchmark in English composition, 40% in algebra, and 26% in biology. Only 22% of the 1.2 million students tested in 2004 met all three benchmarks.

Kobrin (2007) used a model-based method (i.e., logistic regression) to derive an SAT benchmark corresponding to a 65% probability of getting either a 2.7 or 2.0 first-year grade point average (FYGPA). While the model-based procedures used by Kobrin do have the advantage of being empirically validated, there are some potential disadvantages to strictly empirical procedures. Most importantly, such methods may result in cut scores that are unacceptably high or low, and hence violate face validity requirements.

The current study was part of an ongoing effort at the College Board to establish college readiness benchmarks on the SAT and PSAT/NMSQT®, and to provide schools, districts, and states with a comprehensive view of their students' college readiness (Wiley et al., 2010).

Method

This study followed procedures typically used in standard-setting studies to determine cut scores on educational and licensure tests. A standard-setting study is an official research study conducted by an organization that sponsors tests to determine a cut score for the test. To be legally defensible and meet the *Standards for Educational and PsychologicalTesting*

The current study was part of an ongoing effort at the College Board to establish college readiness benchmarks on the SAT and PSAT/NMSQT, and to provide schools, districts, and states with a comprehensive view of their students' college readiness.

(AERA/APA/NCME, 1999), a cut score cannot be arbitrarily determined, and it must be empirically justified. A study is conducted to determine what score best differentiates the classifications of examinees, such as proficient versus not proficient.

In most educational or certification/licensure settings, the traditional process for determining cut scores on a test is to follow standard-setting methods, such as the Angoff or bookmark methods, which require panelists to review the content and student performance on a test, and determine the appropriate score point to set a passing score. In the current study, the decision was made to use a model representing a hybrid between traditional standard-setting and policy-capturing methodologies used in social or organizational psychology (Karren & Barringer, 2002; Kline & Sulsky, 1995). A panel of experts on higher education was convened, but they were not asked to review the test and determine an appropriate passing score. Rather, panelists were asked to draw on their vast experience with student performance in college to determine the most appropriate performance level for successful student performance, using first-year GPA as the available criterion. They were also asked to set a suitable percentage of "college-ready" students who would obtain this GPA. Once these two parameters were determined, empirical procedures were then used to determine the SAT score associated with the parameter values.

The criterion for college success used in this study was first-year college grade point average (FYGPA). This criterion was chosen because the courses that first-year students take are more similar and less variable than at any other year in college, thus minimizing comparability issues that occur with grades. Furthermore, because FYGPA is computed based on grades in many courses taken over students' first year in college, it is a more reliable and representative outcome measure than a single grade in an individual college course. Previous research has also demonstrated that FYGPA is an excellent predictor of eventual graduation (Allen, 1999; Murtaugh, Burns, & Schuster, 1999).

Description of the Panel

Seven highly respected educators and policymakers were invited by the College Board to participate in a day-long meeting in Washington, D.C. on June 16, 2008. The panelists included two vice presidents of enrollment management, one from a large public university and the other from a historically black college; a state commissioner of higher education; the executive director from a state board of education; and three presidents/directors of higher education research centers or education policy organizations.

Data Source: The SAT® Validity Study

The first cohort of students to take the SAT after its last revision in 2005 finished their first year of college in May/June 2007. The College Board contacted colleges and universities across the United States to provide first-year performance data from the fall 2006 entering cohort of first-time students. The sample consisted of individual level data on 195,099 students from 110 colleges and universities across the United States. After limiting the sample to those students with complete data on FYGPA and scores on the latest version of the SAT that includes a writing section, the final sample included 157,983 students. The sample was compared to the population of four-year institutions receiving at least 200 SAT score reports. Table 1 shows the distribution of institutions participating in the validity study by region of the country, selectivity, size, and control, with comparative figures for the population.

Table 1.

Percentage of Institutions by Key Variables: Comparison of Population to Sample

Variable		Population	Sample
Region of U.S.	Midwest	16	15
	Mid-Atlantic	18	24
	New England	13	22
	South	25	11
	Southwest	10	11
	West	18	17
Selectivity	Admits under 50%	20	24
	Admits 50% to 75%	44	54
	Admits over 75%	36	23
Size	Small	18	20
	Medium to large	43	39
	Large	20	21
	Very large	19	20
Control	Public	57	43
	Private	43	57

Note: Percentages may not sum to one hundred due to rounding. With regard to institution size, small = 750 to 1,999 undergraduates; medium to large = 2,000 to 7,499 undergraduates; large = 7,500 to 14,999 undergraduates; and very large = 15,000 or more undergraduates.

The final sample was 54% female and 46% male. The racial/ethnic breakdown of the sample was 67% white/Caucasian, 9% Asian, 7% black, 7% Hispanic, 3% other ethnicity, and less than 1% American Indian. About 7% of the students in the sample did not respond to the SAT Questionnaire item asking for their ethnicity. Nearly all of the students in the sample (90%) reported English as their best language, while approximately 5% reported both English and another language, slightly over 1% reported another language, and 4% did not respond to this question.

On average, the sample for this study performed better on the SAT than the 2006 College-Bound Seniors Cohort. The mean SAT score for the sample was 560 for Critical Reading, 578 for Math, and 554 for Writing, compared to 503, 518, and 497, respectively, in the national cohort. A higher level of performance for the sample compared to the national population of SAT takers was expected, given that all of the students in the sample were enrolled in college. The complete description of the sample, along with the results of the SAT Validity Study, are described in Kobrin, Patterson, Shaw, Mattern, and Barbuti (2008).[2]

Table 2.

Descriptive Statistics on the Total Sample

Predictor	Mean	SD
SAT-CR	560	96.3
SAT-M	578	97.3
SAT-W	554	94.9
FYGPA	2.97	0.71

Note: *N* = 157,983. SAT scores ranged from 200 to 800, and FYGPA ranged from 0 to 4.27.

2. The results of the SAT Validity Study presented in Kobrin et al. (2008) are based on 151,316 students because all students in that study were required to have supplied their self-reported high school grade point average (HSGPA). In this study, HSGPA was not required; therefore, the sample size (157,983) is larger.

The Standard-Setting Process

This study differed from a typical standard-setting study in a few key ways. Typically, standard-setting studies are conducted in order to set cut scores or proficiency levels on tests. The SAT scores identified in this study are not intended to be used as cut scores, and will not be used by institutions in their admission process to accept or reject students. Second, some of the most common forms of standard-setting procedures involve examination of test items or test content. In this study, the panel was *not* asked to examine SAT items, nor were they asked to consider what test content a successful college student could master or not master.

The meeting began with an overview of the goals and purpose of the standard setting, followed by a discussion of current definitions and criteria for college readiness and college success (as summarized in the section "Current Conceptions of College Readiness and College Success"). After this introduction, panelists were asked to provide ratings on what FYGPA should be used to define success in the first year at a four-year college, and the probability that a successful student would attain this FYGPA, or the probability of mastery. (A sample rating form is included in Appendix A.) In addition, the panelists were shown data to illustrate the fact that the Validity Study sample was a more able group than the general population of college applicants, and were asked to take this into consideration when giving their ratings.[3]

It is important to note that in the discussion of criteria for college readiness and college success, many panelists voiced their opinion that the College Board should consider using graduation from college within six years as the measure of success instead of FYGPA. Alternatively, some panelists suggested that the number of credit hours earned should be considered in conjunction with FYGPA to provide a stronger measure of progress toward a college degree. At the time of the meeting, data on graduation and credit hours earned were not available; therefore, panelists were asked to give their ratings based only on FYGPA with the understanding that any benchmarks resulting from these ratings would be further evaluated as more data became available.

The mean ratings for FYGPA and probability of success were used to determine the SAT score associated with those ratings. Using data from the SAT Validity Study, the SAT score that most closely corresponded to the panelists' mean ratings was identified. Data on the percent of students in the Validity Study sample who earned a certain FYGPA were used to help inform the panel's ratings. Table 3 shows the percentage of students in each SAT score range earning FYGPAs of a certain level or higher. For example, of the 16,776 students from the Validity Study sample who scored inclusively between 1400 and 1490 on the three sections of the SAT combined, 74.5% had first-year GPAs of at least 2.30. The information in Table 3 was also made available to the panelists based upon gender and racial/ethnic subgroups, as shown in Appendix B.

3. The panelists were asked to make the assumption that these data, based on the SAT, were representative of all first-year students.

Table 3.

Percentage of Students in the SAT Validity Study Earning Different FYGPA by SAT Score Category

		First-Year GPA								
SAT (CR+M+W)	*n*	2.00	2.30	2.33	2.40	2.50	2.60	2.66	2.70	3.00
600–890	118	72.0	54.2	54.2	50.8	46.6	44.1	39.0	36.4	23.7
900–990	274	68.6	52.2	51.5	46.4	40.5	35.4	32.8	30.7	20.1
1000–1090	842	69.5	52.7	51.2	48.1	42.6	37.1	33.6	31.1	17.7
1100–1190	2,248	73.8	58.5	56.7	52.6	47.8	40.7	37.3	35.5	19.8
1200–1290	5,502	78.7	64.0	62.4	58.0	53.2	47.1	43.7	41.4	24.8
1300–1390	11,001	82.1	69.2	67.8	64.2	59.3	53.2	50.0	47.7	30.6
1400–1490	16,776	85.5	74.5	73.4	70.3	65.7	60.3	57.2	55.1	37.6
1500–1590	21,126	88.4	79.7	78.7	76.0	72.0	67.1	64.0	62.0	44.3
1600–1690	22,752	91.0	84.0	83.2	81.0	77.5	73.2	70.6	68.7	52.2
1700–1790	22,027	93.1	87.5	86.8	85.0	81.9	78.4	76.2	74.5	59.7
1800–1890	19,375	94.9	90.4	89.9	88.3	86.0	83.1	81.4	80.0	66.6
1900–1990	15,245	96.6	93.3	93.0	91.7	90.1	87.9	86.4	85.5	74.1
2000–2090	10,621	97.6	95.3	95.0	94.1	93.0	91.4	90.2	89.4	80.9
2100–2190	6,319	98.5	97.0	96.7	96.3	95.3	94.2	93.3	92.6	86.4
2200–2290	2,836	98.7	97.5	97.4	97.1	96.7	96.0	95.5	95.0	89.2
2300–2400	921	99.1	98.7	98.7	98.5	98.4	97.8	97.6	97.4	94.5

Results

Round 1 Ratings

The panel's first ratings are shown in Figures 1a—1c. The median rating for FYGPA was 2.4, the mean rating was 2.43, and ratings ranged from a low of 2.25 to a high of 2.67, with a standard deviation of 0.14. The median and mean rating for probability levels were 70% and 73%, respectively, ranging from a low of 66% to a high of 80%, with a standard deviation of 5.8%. Based on the average ratings, the SAT score associated with a 73% probability of earning a 2.4 FYGPA ranged from 1490 to 1510. This was determined based on the data in a more detailed version of Table 3, which indicated that approximately 73% of students in the SAT Validity Study who scored 1490, 1500, or 1510 on the SAT earned a 2.40 FYGPA. After the Round 1 results were revealed, the panel was shown the consequences of an SAT benchmark of 1500, both in terms of the percentage of college-bound seniors overall and of the particular subgroups who would meet the benchmark and be labeled "college ready." Overall, in the 2006 cohort, 51.8% of students achieved a benchmark score of 1500. In the Validity Study sample, 76.7% met this benchmark.

Figure 1a.

Frequency distribution of round 1 FYGPA ratings

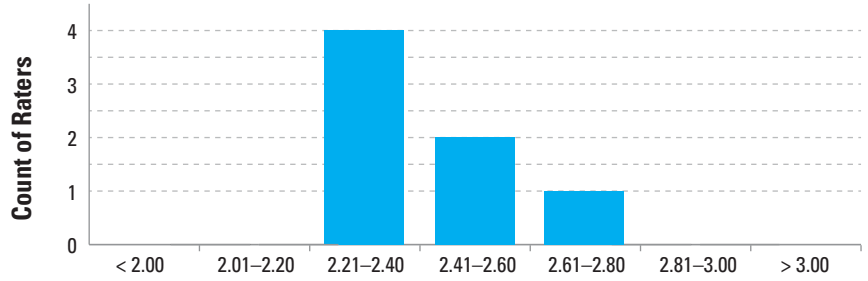


Figure 1b.

Frequency distribution of round 1 probability ratings

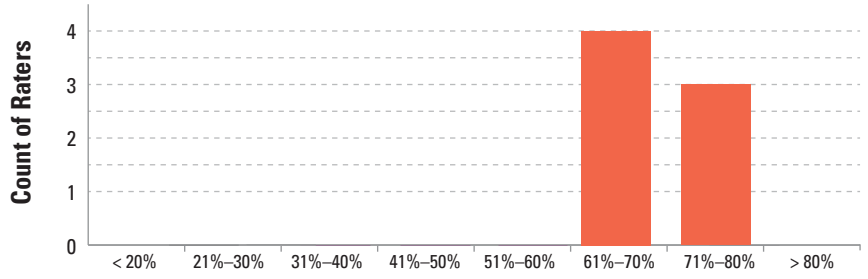
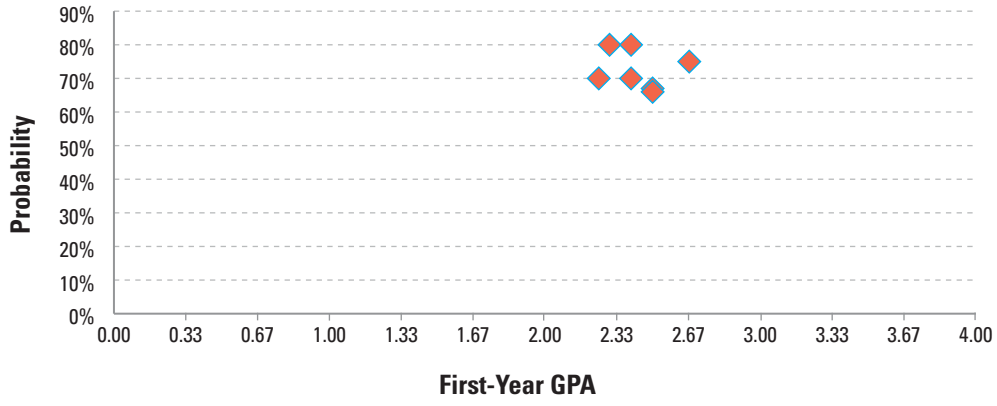


Figure 1c.

Scatterplot of round 1 ratings: FYGPA by probability



Round 2 Ratings

After reviewing the original SAT benchmark value along with its implications on the percentage of students considered college ready overall, and by gender and racial/ethnic groups, the panel provided a second set of ratings. The panel's second ratings are shown

in Figures 2a–2c. The mean rating for FYGPA was 2.62, the median rating was 2.67, and ratings ranged from 2.3 to 3.0, with a standard deviation of 0.25. The discussion after the first set of ratings resulted in some panelists giving higher FYGPA ratings in the second round. The variability among panelists was also higher in the second round. The median and mean ratings for probability level were 70%, ranging from a low of 60% to a high of 75%, with a standard deviation of 6%. Based on the average ratings, the SAT score associated with a 70% probability of earning a 2.6 FYGPA is approximately 1580. Again, as with Round 1, the authors went to the detailed version of Table 3 and identified that about 70% of SAT takers who earned a 1580 composite SAT score were observed to have earned at least a 2.60 FYGPA in college. Table 4 shows the percentage of students meeting the Round 1 and Round 2 benchmarks, both overall and by gender and racial/ethnic subgroups.

The median and mean ratings for probability level were 70%, ranging from a low of 60% to a high of 75%...

Figure 2a.

Frequency distribution of round 2 FYGPA ratings

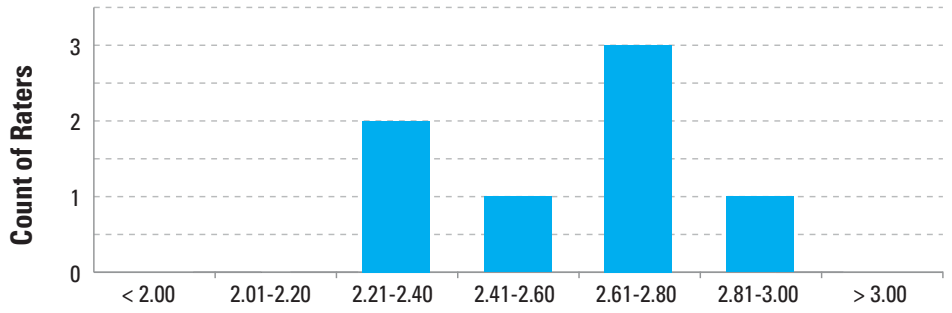


Figure 2b.

Frequency distribution of round 2 probability ratings

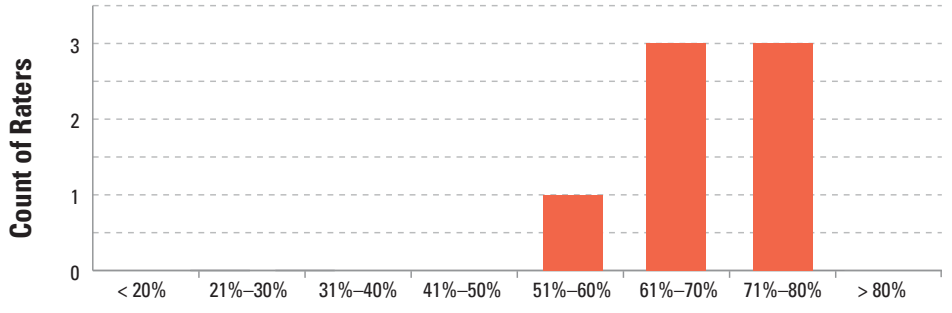


Figure 2c.

Scatterplot of round 2 ratings: FYGPA by probability

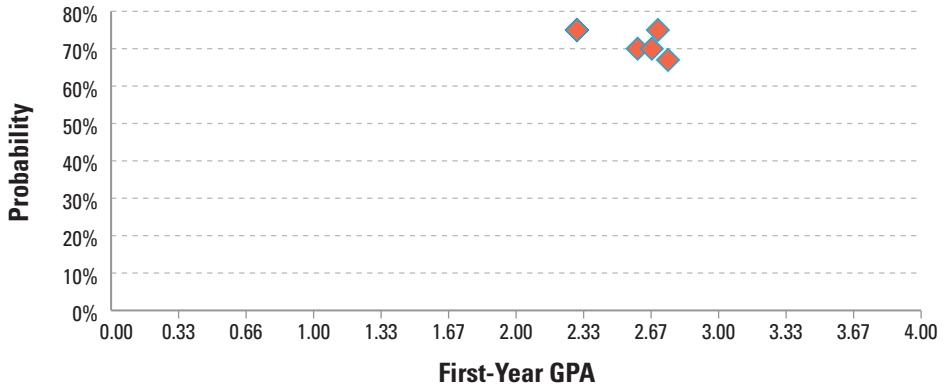


Table 4.

Percentage of Students in the SAT Validity Study Sample and 2006 Cohort Meeting the Benchmarks Overall, and by Gender and Racial/Ethnic Subgroups

Subgroup	SAT Validity Study Sample		2006 College-Bound Seniors Cohort	
	Round 1 Benchmark SAT = 1500	Round 2 Benchmark SAT = 1580	Round 1 Benchmark SAT = 1500	Round 2 Benchmark SAT = 1580
Total Group	76.7	66.2	51.8	41.9
Gender				
Female	74.2	63.1	50.0	40.0
Male	79.7	69.9	53.8	44.0
Race/Ethnicity				
American Indian	70.2	58.4	43.9	33.2
African American	50.4	37.1	20.9	14.1
Asian American	81.3	72.3	60.7	51.8
Hispanic	61.6	49.6	30.8	22.2
White	80.2	69.7	61.0	49.8
Other	74.8	64.8	49.2	39.5

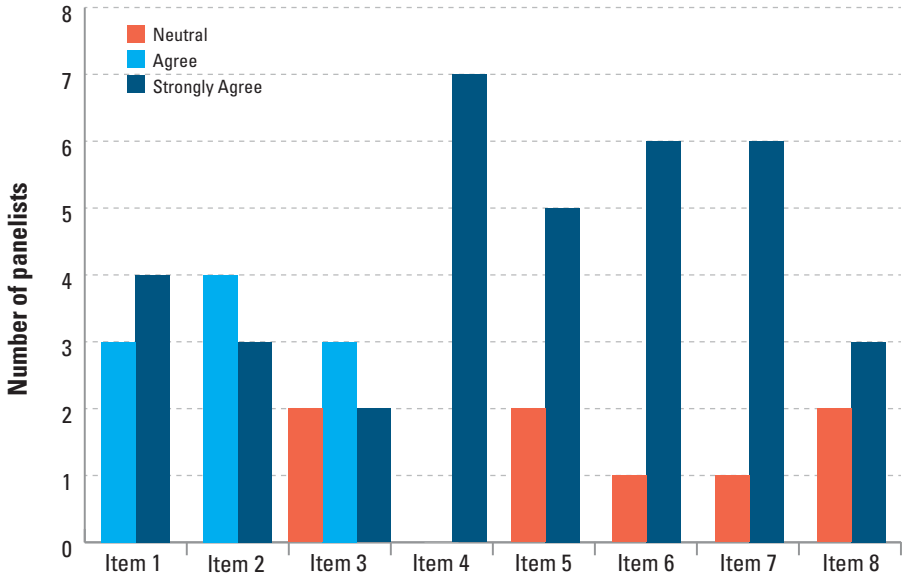
Standard-Setting Evaluation

At the conclusion of the meeting, the panelists were asked to complete a survey to evaluate the meeting and the process for determining benchmarks on the SAT. (A copy of the evaluation survey is included in Appendix C.) Overall, the panelists were comfortable with the task assigned to them, and with the standard-setting process. As shown in Figures 3a and 3b, most of the panelists either agreed or strongly agreed with all statements. All seven panelists strongly agreed that they were comfortable contributing to the discussions. The aspects of the standard setting garnering the highest ratings were with regard to the role of the facilitator, the conduciveness of the meeting facilities to accomplishing the tasks, and the timing and pace of the meeting. Three panelists felt that the SAT score established at the meeting was "about right," one panelist felt it was too high, and three panelists did not respond to this question. None of the panelists disagreed, or strongly disagreed, with any of the survey items.

The panelists were asked to indicate what sources of information they used in their ratings, and their level of reliance on each source of information (using the options heavily, moderately, slightly, or not at all). Figure 3c shows that the group discussion was the most heavily used source of information, followed by the panelists' prior knowledge of other college success or benchmark studies.

Figure 3a.

Evaluation survey results, part I (items 1–8)

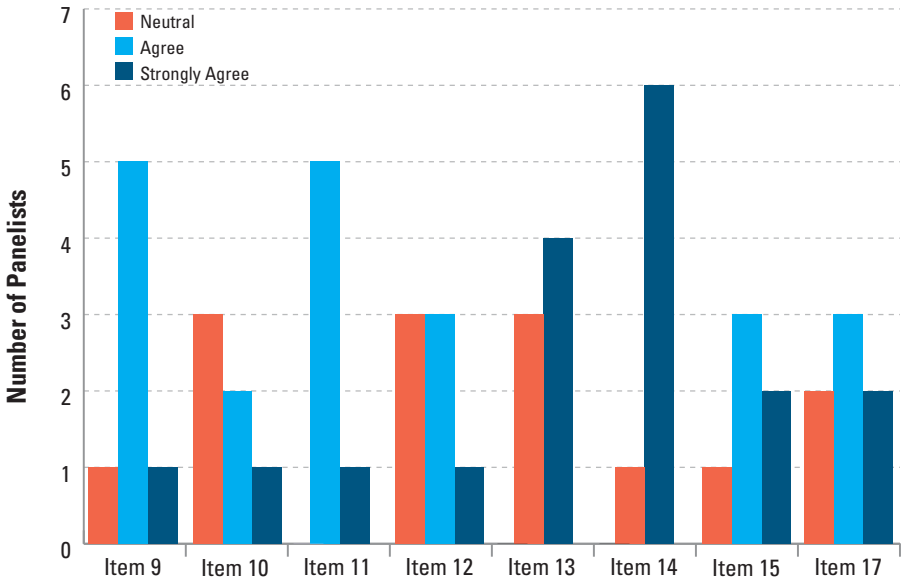


Item 1: The introduction provided a clear overview of the purpose of standard setting for determining benchmarks for the SAT.
Item 2: The introduction answered questions I had about the purpose of determining benchmarks on the SAT.
Item 3: I had a good understanding of my role in this standard-setting meeting.
Item 4: I felt comfortable contributing to the discussions.
Item 5: The facilitator helped me understand the standard-setting process.
Item 6: The meeting facilities and materials were conducive to accomplishing the tasks.
Item 7: The timing and pace of the meeting were appropriate.
Item 8: I was comfortable with the criterion/definitions of college success discussed at the meeting.

Overall, the panelists were comfortable with the task assigned to them, and with the standard-setting process. As shown in Figures 3a and 3b, most of the panelists either agreed or strongly agreed with all statements.

Figure 3b.

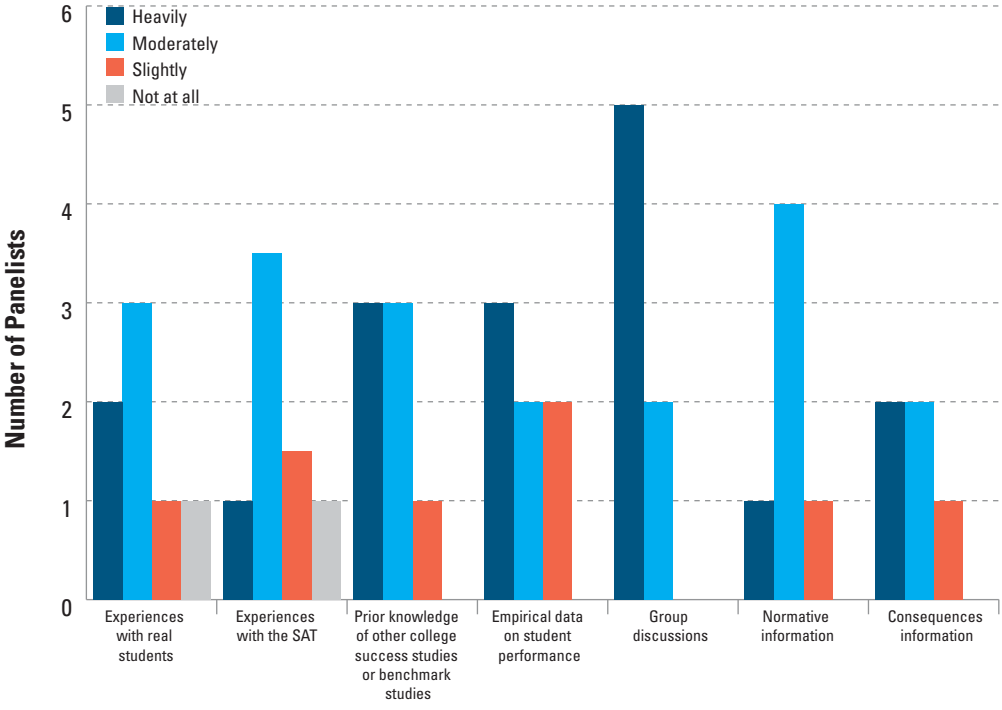
Evaluation survey results, part I (items 9–15 and 17)



Item 9: The data source used to determine the SAT benchmarks (i.e., SAT Reasoning Test™ Validity Study) was appropriate.
Item 10: The data presented provided enough information to complete my ratings.
Item 11: I understood how to complete my ratings.
Item 12: I was confident in my ratings.
Item 13: Overall, the meeting allowed me to use my experience and expertise to recommend benchmarks for the SAT.
Item 14: Overall, the facilitator helped to ensure that everyone was able to contribute to the group discussions and that no one dominated the discussions.
Item 15: Overall, I was able to understand and use the information provided (e.g., other participants' ratings, impact data).
Item 17: The SAT benchmark will assist states and secondary school systems to determine the college readiness of students.

Figure 3c.

Evaluation survey results, part II



Discussion

As Jaeger (1976, cited in Koffler, 1980) aptly summarized over 30 years ago, "No amount of data collection, data analysis and model building can replace the ultimate judgmental act of deciding which performances are meritorious or acceptable and which are unacceptable or inadequate" (page 22). The work completed in this study to establish a criterion for college success provides support for the identification of an SAT benchmark that can be used by states and school districts to gauge the college readiness of their students. In 2011, the College Board introduced its college- and career-readiness benchmark, which is a combined score on the SAT critical reading, mathematics, and writing section of 1550. This benchmark indicates a 65% likelihood of achieving a B- average or higher during the first year of college (College Board, 2011). The choice of this benchmark score was based in part on the results of the standard-setting study described in this report.

Previous efforts to establish indicators of college readiness have taken different methodological approaches, and have focused on different criteria of college success. The expert panel initially recommended using college graduation within six years as an indicator of success, but given that these data were not available, they agreed that FYGPA could serve as a reasonable measure of college success. Furthermore, research has established a strong link between FYGPA and retention (Allen, 1999; Murtaugh et al., 1999). Murtaugh, Burns, and Schuster found that the likelihood of being retained for four years increased from 33% for students with the lowest first-quarter GPAs (0.0–2.0) to 76% for students with the highest first-quarter GPAs (3.3–4.0). Therefore, FYGPA may be considered a good indicator of college success, and a precursor of persistence and graduation.

Nonetheless, while the SAT is a strong predictor of college success, it is acknowledged that many factors other than academic preparation are essential to successfully complete a college degree. For example, one important quality of students who succeed in college is the ability to navigate through the college system. Roderick, Nagaoka, Coca, and Moeller (2008) conducted a study to understand the college search and application behaviors among Chicago Public School (CPS) students, as well as the extent to which high schools can create environments that support students in thoroughly engaging in this process. Among CPS students who aspired to attain a four-year degree, only 41% took the steps necessary in their senior year to apply to, and enroll in, a four-year college. An additional 9% of students managed to enroll in a four-year college without following the standard steps, for a total of 50% of all CPS students who aspired to a four-year degree. It is recognized that the ability to navigate the college system is a very important component for college success; however, this will not be incorporated into the College Board's benchmarks because at present there is no test or instrument that captures this trait.

The standard-setting panel stressed that many other factors are associated with a student's ability to succeed in college, including the student's financial need, number of hours worked, and whether the student is the first generation in his/her family attending college. These factors should be considered along with academic preparation when determining what constitutes college success.

References

ACT. (2004). *Crisis at the core: Preparing all students for college and work*. Iowa: ACT, Inc. Retrieved from http://www.act.org/research/policymakers/pdf/crisis_report.pdf

Achieve, Inc., The Education Trust, & Thomas B. Fordham Foundation. (2004). *The American diploma project: Ready or not: Creating a high school diploma that counts*. Washington, DC: Achieve, Inc. Retrieved from http://www.achieve.org/files/ADPreport_7.pdf

Allen, D. (1999). Desire to finish college: An empirical link between motivation and persistence.*Research in Higher Education*, *40*, 461–485.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Berkner, L., & Chavez, L. (1997). *Access to postsecondary education for the 1992 high school graduates* (NCES 98-105). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs98/98105.pdf

College Board. (2011). 43% of 2011 college-bound seniors met SAT college and career readiness benchmark [Press release]. Retrieved from http://media.collegeboard.com/pdf/cbs_2011_nat_release_091411.pdf

Conley, D. T. (2007). *Toward a more comprehensive conception of college readiness*. Eugene, OR: Educational Policy Improvement Center. Retrieved from http://www.collegiatedirections.org/2007_Gates_CollegeReadinessPaper.pdf

Corwin, Z. B., & Tierney, W. G. (2007). *Getting there – and beyond: Building a culture of college-going in high schools*. Los Angeles, CA: University of Southern California Center for Higher Education Policy Analysis. Retrieved from http://www.usc.edu/dept/chepa/working/Getting%20There%20FINAL.pdf

Greene, J. P., & Winters, M. A. (2005, February). *Public high school graduation and college-readiness rates: 1991–2002* (Education Working Paper No. 8). Retrieved from http://www.manhattan-institute.org/html/ewp_08.htm

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. *Florida Journal of Educational Research*, *18*, 22–27.

Karren, R.J., & Barringer, M.W. (2002). A review and analysis of the policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods*, *5*(4), 337–361.

Kirst, M. W., & Venezia, A. (2006). *Improving college readiness and success for all students: A joint responsibility between K–12 and postsecondary education* (Issue Brief for the Secretary of Education's Commission on the Future of Higher Education). Retrieved April 21, 2008, from http://www.ed.gov/about/bdscomm/list/hiedfuture/reports/kirst-venezia.pdf

Kline, T.J.B., & Sulsky, L.M. (1995). A policy-capturing approach to individual decision-making: A demonstration using professors' judgements of the acceptability of psychology graduate school applicants. *Canadian Journal of Behavioural Science*, *27*(4), 393–404.

Kobrin, J. (2007). *Determining SAT benchmarks for college readiness* (College Board Research Note RN-30). New York: The College Board. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/RN-30.pdf

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (College Board Research Report No. 2008-5). New York: The College Board. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/08-1718_RDRR_081017_Web.pdf

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement*, *17*(3), 167–178.

Lord, J. M. (2003). *ACT and SAT scores in the South: The challenge to lead* (SREB ED 476418). Atlanta, GA: Southern Regional Education Board College Readiness Series. Retrieved from http://publications.sreb.org/2003/03E47_ACT-SAT_2003_.pdf

McNeil, M. (2008, March 12). Benchmarks momentum on increase: Governors' group, state chiefs eyeing international yardsticks. *Education Week*, *27*(27), 12–13.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*, 355–371.

New England Board of Higher Education (2006). *College ready New England: Leaders goals and policy strategies*. White paper. Boston, MA: Author.

Roderick, M., Nagaoka, J., Coca, V., & Moeller, E. (2008). *From high school to the future: Potholes on the road to college*. Retrieved June 25, 2008, from the University of Chicago, Consortium on Chicago School Research website: http://ccsr.uchicago.edu/downloads/1835ccsr_potholes_summary.pdf

Southern Regional Education Board. (2002). *Student readiness for college: Connecting state policies*. Atlanta, GA: Author. Retrieved from http://publications.sreb.org/2002/02E06_Student_Readiness2002.pdf

Twing, J.S., Miller, G.E., & Meyers, J. L. (2008). *TAKS Higher Education Readiness Component (HERC) contrasting groups study* (Pearson Research Bulletin). Pearson Educational Measurement.

Wiley, A., Wyatt J., & Camara, W. J. (2010). *The development of a multidimensional college readiness index* (College Board Research Report No. 2010-3). New York: The College Board. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/10b_3110_CollegeReadiness_RR_WEB_110315.pdf

Von Secker, C. (2009). *Closing the gap: Seven keys to college readiness for students of all races/ethnicities* (Accountability update). Montgomery County Public Schools, Applied Research Unit. Retrieved from http://205.222.0.20/info/keys/documents/research.pdf

Appendix A: Sample Rating Form

SAT Benchmarks Standard-Setting Study

June 16, 2008

Panel Member Information

Name:_____

Title: _____

Affiliation:_____

Rating Sheet for Round 1

First-Year Grade Point Average (FYGPA)

Note: First-Year Grade Point Average should be expressed on a 0.00 to 4.00 scale. In many institutions, letter grades map to this scale such that 4.00 = A; 3.00 = B; 2.00 = C; and 1.00 = D.

Minimum Probability of Mastery

Note: Probability level should be expressed in terms of percent on a 0% to 100% scale.

Appendix B

Table B1.

Percentage of Students in the SAT Validity Study Earning Different FYGPA by SAT Score Category, and by Gender and Ethnic Subgroups

	Females					Males				
SAT Score	*n*	2.00	2.30	2.60	3.00	*n*	2.00	2.30	2.60	3.00
600–890	60	66.7	50.0	43.3	21.7	58	77.6	58.6	44.8	25.9
900–990	146	66.4	51.4	32.9	17.8	128	71.1	53.1	38.3	22.7
1000–1090	502	71.3	54.8	38.8	18.7	340	66.8	49.7	34.4	16.2
1100–1190	1,383	76.1	61.1	43.3	21.3	865	70.1	54.3	36.5	17.5
1200–1290	3,387	81.1	67.0	51.0	28.0	2,115	74.9	59.3	40.9	19.7
1300–1390	6,571	84.8	73.0	57.8	34.4	4,430	78.1	63.5	46.4	25.0
1400–1490	9,862	88.3	78.8	65.8	42.9	6,914	81.5	68.4	52.5	30.0
1500–1590	11,943	91.3	84.2	72.9	50.6	9,183	84.5	73.9	59.6	36.1
1600–1690	12,398	94.0	88.6	79.7	59.7	10,354	87.5	78.5	65.5	43.2
1700–1790	11,513	96.0	92.1	84.6	67.3	10,514	89.9	82.4	71.6	51.3
1800–1890	9,768	97.2	94.2	88.6	74.1	9,607	92.5	86.6	77.5	58.9
1900–1990	7,446	98.1	96.3	92.7	81.1	7,799	95.1	90.6	83.4	67.5
2000–2090	5,115	99.0	97.6	94.8	87.1	5,506	96.3	93.2	88.2	75.2
2100–2190	2,999	99.2	98.4	96.6	91.1	3,320	97.9	95.6	92.0	82.1
2200–2290	1,308	99.2	98.5	97.4	92.7	1,528	98.3	96.6	94.8	86.3
2300–2400	393	99.7	99.7	99.0	96.7	528	98.7	97.9	97.0	92.8
	American Indian					African American				
SAT Score	*n*	2.00	2.30	2.60	3.00	n	2.00	2.30	2.60	3.00
600–890	n/r	n/r	n/r	n/r	n/r	38	57.9	39.5	28.9	13.2
900–990	n/r	n/r	n/r	n/r	n/r	95	63.2	44.2	30.5	15.8
1000–1090	n/r	n/r	n/r	n/r	n/r	220	65.9	43.2	33.2	16.4
1100–1190	17	52.9	47.1	35.3	23.5	506	68.8	54.7	35.0	13.8
1200–1290	40	77.5	65.0	60.0	25.0	997	75.4	56.7	37.8	18.3
1300–1390	70	75.7	65.7	42.9	21.4	1,520	76.1	61.9	42.7	21.4
1400–1490	109	79.8	69.7	51.4	29.4	1,829	80.4	67.9	52.9	29.9
1500–1590	116	81.0	70.7	60.3	37.1	1,725	84.3	73.4	57.6	34.3
1600–1690	148	86.5	78.4	65.5	47.3	1,397	90.6	80.7	68.9	46.0
1700–1790	100	92.0	85.0	79.0	48.0	981	90.9	82.1	70.5	49.1
1800–1890	88	89.8	85.2	79.5	61.4	604	92.2	84.1	74.2	53.0
1900–1990	70	92.9	85.7	81.4	52.9	334	95.2	90.1	82.3	68.0
2000–2090	24	100.0	91.7	87.5	70.8	173	96.0	89.0	82.1	69.4
2100–2190	15	100.0	100.0	93.3	93.3	49	95.9	91.8	85.7	69.4
2200–2290	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r
2300–2400	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r	n/r

Note: Values based on fewer than 15 students are not reported.

Appendix B, *continued*

	Asian American					Hispanic				
SAT Score	*n*	2.00	2.30	2.60	3.00	*n*	2.00	2.30	2.60	3.00
600–890	n/r	n/r	n/r	n/r	n/r	20	75.0	55.0	45.0	20.0
900–990	24	95.8	75.0	58.3	37.5	34	58.8	50.0	35.3	17.6
1000–1090	63	79.4	74.6	60.3	30.2	172	62.8	45.3	32.6	16.3
1100–1190	155	76.8	66.5	49.7	31.6	339	67.6	52.5	37.5	17.1
1200–1290	419	85.4	75.2	58.9	33.4	719	74.1	58.1	41.6	20.3
1300–1390	799	87.5	75.3	60.6	35.7	1,302	77.0	62.0	45.8	25.9
1400–1490	1,267	90.2	78.5	64.2	41.7	1,566	79.4	66.1	51.7	31.8
1500–1590	1,653	89.1	79.4	67.8	44.7	1,644	85.8	76.2	62.3	39.9
1600–1690	1,827	90.8	84.9	74.2	54.2	1,539	88.9	80.4	67.9	45.2
1700–1790	1,932	93.1	87.1	78.0	58.3	1,303	92.0	85.4	74.2	51.8
1800–1890	1,918	94.4	89.4	81.4	64.6	980	94.7	89.3	81.0	60.4
1900–1990	1,682	96.1	92.6	85.9	71.4	605	95.9	92.1	83.1	65.5
2000–2090	1,352	97.1	94.5	89.2	78.3	351	96.9	92.6	86.0	73.2
2100–2190	867	98.6	96.1	92.7	83.9	159	98.7	95.0	91.8	81.8
2200–2290	475	97.7	95.8	93.7	86.1	58	98.3	94.8	94.8	84.5
2300–2400	192	99.0	97.9	96.4	93.2	n/r	n/r	n/r	n/r	n/r
	White					Other				
SAT Score	*n*	2.00	2.30	2.60	3.00	*n*	2.00	2.30	2.60	3.00
600–890	25	68.0	60.0	48.0	16.0	n/r	n/r	n/r	n/r	n/r
900–990	75	76.0	57.3	33.3	16.0	n/r	n/r	n/r	n/r	n/r
1000–1090	259	71.0	55.6	36.3	15.4	29	82.8	75.9	48.3	27.6
1100–1190	927	78.5	62.4	46.2	22.8	108	77.8	58.3	35.2	19.4
1200–1290	2,773	80.6	66.7	50.5	27.6	181	79.6	61.3	42.0	20.4
1300–1390	6,308	83.8	71.6	56.3	32.9	355	83.1	67.0	53.2	32.1
1400–1490	10,549	86.9	76.6	62.8	39.4	482	85.3	72.8	60.6	38.8
1500–1590	14,215	89.1	80.9	68.6	46.1	598	88.8	82.1	69.9	46.3
1600–1690	15,943	91.5	84.8	74.2	53.3	636	89.0	82.4	70.4	50.3
1700–1790	15,770	93.4	88.1	79.3	61.4	608	92.8	87.5	79.4	59.4
1800–1890	13,892	95.2	91.0	84.1	68.2	560	93.9	90.0	82.3	67.3
1900–1990	10,835	96.8	93.7	88.7	75.8	468	96.4	94.0	88.2	73.7
2000–2090	7,438	97.7	95.6	92.1	82.1	297	96.0	92.9	88.2	76.8
2100–2190	4,378	98.5	97.2	94.9	87.4	175	98.3	97.1	94.9	88.0
2200–2290	1,855	99.2	98.1	96.6	90.2	88	96.6	94.3	94.3	87.5
2300–2400	533	98.9	98.7	98.1	94.6	29	100.0	100.0	100.0	96.6

Note: Values based on fewer than 15 students are not reported.

Appendix C: Standard-Setting Evaluation Survey

Part 1

Directions: Please check one box for each of the following statements by placing an "X" in the box corresponding to your opinion. If you have any additional comments, please write them in the space provided at the end of this form.

KEY: SD = Strongly Disagree; D = Disagree; N = Neutral; A = Agree; SA = Strongly Agree

	Statement	SD	D	N	A	SA
1	The introduction provided a clear overview of the purpose of standard-setting for determining benchmarks for the SAT.					
2	The introduction answered questions I had about the purpose of determining benchmarks on the SAT.					
3	I had a good understanding of my role in this standard-setting meeting.					
4	I felt comfortable contributing to the discussions.					
5	The facilitator helped me understand the standard-setting process.					
6	The meeting facilities and materials were conducive to accomplishing the tasks.					
7	The timing and pace of the meeting were appropriate.					
8	I was comfortable with the criterion/definitions of college success discussed at the meeting.					
9	The data source used to determine the SAT benchmarks (i.e., SAT Reasoning Test Validity Study) was appropriate.					
10	The data presented provided enough information to complete my ratings.					
11	I understood how to complete my ratings.					
12	I was confident in my ratings.					
13	Overall, the meeting allowed me to use my experience and expertise to recommend benchmarks for the SAT.					
14	Overall, the facilitator helped to ensure that everyone was able to contribute to the group discussions and that no one dominated the discussions.					
15	Overall, I was able to understand and use the information provided (e.g., other participants' ratings, impact data).					
16	The final group-recommended benchmark is:	_____ too high _____ too low _____ about right				
17	The SAT benchmark will assist states and secondary school systems to determine the college readiness of students.					
18	I have concerns about the SAT benchmarks and their potential use or misuse. (If you answer A or SA, please explain below.)					
	Please give additional comments or concerns about the SAT benchmarks, the standard-setting meeting, the College Board's use or dissemination of the benchmarks, etc..					

Appendix C, *continued*

Part 2

Directions: The list below contains all of the sources of information that were available for generating your ratings during the standard-setting process. For each question below, first place an "X" in one box following each source of information to indicate how much you relied on that source of information to make your judgments. Please mark one box in each row.

Second, consider which source of information you relied upon most, and which you relied upon least, to make your judgments. Place one plus sign (+) in the column at the far right to indicate the one source you relied upon most, and one minus sign (-) to indicate the one source you relied upon least.

Sources of Information	Level of Reliance on Information				
	Heavily	Moderately	Slightly	Not at all	+/–
My own experiences, knowledge, and/or opinions regarding real students					
My own experiences, knowledge, and/or opinions regarding the SAT					
Prior knowledge of other college success studies or benchmark studies					
The empirical data on student performance					
The group discussions					
The normative information (i.e., the ratings of other participants)					
The consequences information (i.e., impact data)					
Other (specify):_____ _____					

The Research & Development department actively supports the College Board's mission by:

- Providing data-based solutions to important educational problems and questions
- Applying scientific procedures and research to inform our work
- Designing and evaluating improvements to current assessments and developing new assessments as well as educational tools to ensure the highest technical standards
- Analyzing and resolving critical issues for all programs, including AP®, SAT®, PSAT/NMSQT®
- Developing standards and conducting college and career readiness alignment studies
- Publishing findings and presenting our work at key scientific and education conferences
- Generating new knowledge and forward-thinking ideas with a highly trained and credentialed staff

Our work focuses on the following areas

Admission	Measurement
Alignment	Research
Evaluation	Trends
Fairness	Validity

Follow us online: collegeboard.org/research

CollegeBoard

CollegeBoard

12b-5030