Khan Academy

Use of Khan Academy Official SAT Practice and SAT Achievement: An Observational Study

Technical Report

First released: 17 August 2020

For questions related to the study, please contact efficacy@khanacademy.org.
For press related questions, please contact press@khanacademy.org.
Click here to access an ADA accessible version of this report.

Acknowledgments
We want to thank our methodological reviewers, Derek Briggs and Laura O'Dwyer, for their detailed reviews and feedback on this report, and to thank our external panel of reviewers, Richard Bowman, Judie Cherenfant, Christopher Dupuis, Erica Ramirez Horvath, Mercedes Pour, Simone Rahotep, and Ferdinand Wipachit, for their guidance and feedback on this report.

Disclaimer
All statements and conclusions, unless specifically attributed to another source, are those of the authors and do not necessarily reflect the policies or positions of the other organizations or references noted in this report. For questions or comments about this study, please contact **efficacy@khanacademy.org**.

The Khan Academy Efficacy & Research team releases our technical reports as working papers to quickly share the results of our latest studies. As such, these working papers have not undergone blind peer review, but they were reviewed externally by experts before their release. We release our results early as a way to obtain feedback and discussion from the research community before submitting them for publication in a peer-reviewed journal.

Authors
Kodi Weatherholtz, Phillip Grimaldi,* Catherine Hicks,* Kelli Millwood Hill
Khan Academy
* *these authors contributed equally to this work*

Cassie Freeman, Bercem Akbayin-Sahin, Crystal Coker, Jennifer Ma, Lara Henneman
College Board

Suggested Citation
Weatherholtz, K., Grimaldi, P., Hicks, C., Hill, K.M., Freeman, C., Akbayin-Sahin, B., Coker, C., Ma, J., & Henneman, L. (2020). *Use of Khan Academy Official SAT Practice and SAT Achievement: An Observational Study*. Mountain View, CA: Khan Academy.

Contents

Executive Summary

At a time when more education is taking place online than ever before, students and educators need targeted guidance to make the most of their instructional time. SAT® preparation and the journey to college is no exception. Since 2015, Official SAT Practice (OSP) on Khan Academy® has provided free, personalized practice in an online format to help all students build their skills and prepare for the SAT. More than 10 million students have used OSP since the launch.

In this report, Khan Academy and College Board jointly analyze OSP usage by more than half a million students in the class of 2019 between their PSAT/NMSQT® and their first SAT in order to associate the use of OSP with their SAT performance. This builds on our prior work that showed a positive association between OSP use and higher scores. While our previous work focused on practice between the PSAT/NMSQT and the last SAT, we are now focusing our interval to the first SAT in order to better isolate the impact of OSP, particularly from the intervening SAT assessments a student may have taken.

In our 2017 study with approximately 250,000 "early adopter" students, we observed a 90-point score increase overall (from PSAT/NMSQT to the last SAT) for students using OSP. After removing the typical growth between PSAT/NMSQT and the last SAT to examine the impact of OSP directly, the specific *added* growth from spending six to eight hours practicing on OSP was 30 additional points on their last SAT compared to students who did not use OSP. In this current study, with wider-spread adoption of OSP, we found similar results for the PSAT/NMSQT to the first SAT interval. Specifically, we find that students who spent six or more hours of practice on OSP scored an additional 21 points higher on their first SAT than students who did not use OSP, but 39 points higher when students used at least one of the best practice behaviors (described below). These findings hold true across student demographics, including gender, race/ethnicity, and level of parental education.

This report looks at how OSP works for a broader population of students and outlines new insights on three best practices of meaningful OSP use, which can optimize student time on the platform. OSP best practice behaviors are actions any student can take during their practice that are associated with greater scores on the SAT. These behaviors were operationalized based on how the product is designed, the data elements available, and prior research concerning the effectiveness of test preparation strategies. We also find that the best practice behaviors are correlated, but students show selective strategies of how they engage with OSP. The three best practice behaviors include:

- **Leveling up skills:** As students progress through OSP material, they can achieve new levels in the skills practiced. Overall, leveling up provides a signal that students are consistently advancing in the content tested on the SAT, and is a marker for learning progress on OSP. This best practice behavior also helps students learn how to monitor their progress.
- **Taking a full-length practice test**: Taking a full-length practice test simulates the real test experience and helps students see what they do know and don't know. Eight full-length online practice tests, which can be taken in one sitting or over time, are available on OSP.
- **Following personalized skill recommendations**: OSP provides personalized skill recommendations based on a student's previous scores and performance on any PSAT-related test or SAT assessment or through mini-diagnostic quizzes. Following the personalized skill recommendations helps a student learn how to stay focused when they study and work on areas where they need the most help.

More time spent on OSP is associated with higher scores on the SAT. However, time spent is not enough. Best practice behaviors can help guide students and ensure that the time spent on practice is productive.

Although engaging in best practice behaviors was associated with improved performance, students varied considerably in their adoption. Indeed, the majority of students in our sample unfortunately did not engage in any of the best practice behaviors (roughly 8% of students in our sample spend six or more hours and complete a best practice on OSP). Differences in student background, household, and demographics were associated with the likelihood to engage in a best practice behavior. Although these characteristics are associated with student behavior, they are likely rough indicators of other factors, including different educational environments, that may impact how students practice. It is also important to note that these between-group differences were small and did not result in meaningful differences between groups in terms of their benefits from using OSP. These results signal that more work is needed to point students to best practice behaviors and to motivate their usage across the platform. In coming years, College Board and Khan Academy will work diligently with our partners across the country through programmatic supports and platform refinements to ensure all students can follow these best practices.

Although the data associating best practices with score increases are promising, we need more research on implementation to ensure that when best practices are used more broadly, the associations remain as strong. Further research will help to build our understanding of student progress; any differences in adoption of best practice behaviors; and how supports such as school-day implementation and educator tools can help keep all students engaged and on track. In the ever-evolving educational landscape, it is our hope that sharing and continuing this research on the evidence for the use of best practice behaviors can make a difference for the millions of students who use the platform on their path to college, and that students can make the most effective use of their time on OSP and, ultimately, be successful in their efforts.

Introduction

The mission of Khan Academy is to provide a free, world-class online education to anyone, anywhere. It is available in 40 different languages and 18 million people use Khan Academy each month. In 2019, the Khan Academy website included 429 courses, 4,347 articles, 74,507 problems, and 13,327 videos for learners worldwide. These resources span K–16 subjects, including grade-specific K–12 courses in math, science and engineering, computing, arts and humanities, economics and finance, test prep, and college and careers.

In spring 2014, Khan Academy entered a partnership with College Board, the administrator of the SAT®, to provide free SAT practice. By summer 2015, Khan Academy released Official SAT Practice (OSP). OSP creates a personalized plan that will help each student prepare for the SAT. Included are thousands of interactive questions with instant feedback, video lessons, eight full-length practice tests, and more. To receive a personalized practice plan, students can either take a series of diagnostic quizzes or link their College Board and Khan Academy accounts. Additional details on the OSP features are discussed later in the Product Features section.

The purpose of this report is to describe findings from a large-scale analysis—conducted jointly by researchers at Khan Academy and College Board—concerning the relationship between students' use of Khan Academy OSP and their SAT achievement. We investigate the following research questions:

Descriptive findings of OSP usage patterns

- 1a. What does student usage on OSP look like?
- 1b. When are students using OSP?
- 1c. How are students engaging in best practices?

Associations between OSP and SAT performance

- 2a. Does time spent using OSP relate to SAT achievement?
- 2b. Do all students benefit equally from their time spent on OSP?
- 2c. Are best practice behaviors associated with improved SAT performance?
- 2d. Are certain students more likely to engage in best practices?

In addressing these questions, we have the following goals: (i) to study the association between OSP usage and SAT scores; (ii) to illuminate particular types of behaviors on OSP that vary in their association with SAT performance and to identify whether certain groups of students are more/less likely to benefit from OSP.

Background

Each year, millions of students take the SAT for college admissions and scholarship opportunities (College Board, 2019a). To prepare for the SAT, students face a wide array of products and services that vary in cost, personalization, and quality. Free options include released exams, library books, courses offered by nonprofits, courses offered at school, promotional sessions by test prep companies, and OSP on Khan Academy. In addition to free offerings, some families can afford paid access to online resources, test prep books, courses, and individual tutors. With an ever-growing industry, the cost of courses and tutoring can be substantial (Buchmann et al., 2010; U.S. News and World Report, 2020). Previous studies on test prep have focused on products (e.g., courses) and services (e.g., tutoring), as well as dosage (e.g., number of sessions, hours). Generally, these studies found that test prep had a positive impact on students' test scores, although the estimated impact varied in magnitude and was typically within the margin of error of the assessment (Appelrouth et al., 2018; Briggs, 2009; Buchmann et al., 2019, Moore et al., 2019).

Prior Research on Test Preparation Effectiveness

Previous research on test preparation effectiveness has focused on the attributes of students who have access to preparation as well as the impact of that preparation on their ACT and SAT outcomes. Differences in who has access to various types of ACT and SAT preparation have been associated with family income, race of the student, parental education, and high school environment. Students from higher income families are consistently more likely to enroll in paid in-person coaching classes and private tutoring. However, the role of race and parental education is more unclear. Recent studies found that East Asian and Black students were more likely than their White peers to take paid coaching classes and students with higher parental education levels were more likely to take paid coaching classes, but not private tutoring (Buchmann et al., 2010; NACAC, 2008; Briggs, 2001; Byun & Park, 2012; Park & Beck, 2015). Students also generally engage in more than one type of test preparation, which makes it difficult to estimate the impact of any one particular preparation type. In a survey of spring 2018 SAT test takers, 71% of respondents said they used at least two approaches to practice, with the combination of Official SAT Practice and test prep books being the most frequent; 18% of all respondents used this combination (College Board 2018a). In a survey of recent ACT retesters, 34% reported engaging in four or more test preparation activities (ranging from free online programs to paid tutoring), with low-income and minority students engaging in fewer test preparation activities (Moore et al., 2019).

The impacts of preparation vary across interventions and studies, with coaching associated with about 18 to 33 additional points on the math portion of the SAT and about 8 to 24 additional points on the verbal portion (Briggs, 2009; Montgomery & Lilly, 2012). Coaching is broadly defined as systematic test preparation involving content review, question drills, specific test-taking strategies (e.g., eliminating answers, active reading, plugging in numbers), and general knowledge about the structure of the exam, all with the aim of increasing test scores (Briggs, 2002). These impacts of coaching, generally within .25 standard deviations, are modest when compared to claims made by test prep companies (Briggs, 2009; U.S. News and World Report, 2020). Impacts associated with preparation may matter practically, especially when colleges use score cut-points in admissions decisions (Briggs, 2009).

Across preparation modalities, the dosage of preparation has been an area of interest. One analysis found that each additional hour of tutoring was associated with an increase of 2.34 SAT points (Appelrouth et al., 2015) and another analysis found that six to eight hours of OSP usage was associated with an additional 30-point score increase from the PSAT/NMSQT® to students' last SAT (College Board, 2018b). More recently, Moore and colleagues reported that 11 hours or more of tutoring between a first and second ACT was associated with a 0.60-point increase, on a scale of 1 to 36, compared to students who did not have tutoring (Moore et al., 2019).

With the exception of retesting, few studies have closely examined the impact of particular preparation activities (practice problems vs. lessons, mini-sections, timed tests) and cadence on SAT outcomes (Moore et al., 2019; Appelrouth et al., 2018). Evidence from the Appelrouth et al. (2015, 2018) studies of students in tutoring suggests that starting spaced preparation early (before June of a student's junior year), timed practice tests, multiple official SAT tests, and sufficient instructional time all have positive associations with SAT outcomes, ranging from 20 additional points for an official SAT practice test to 42 additional points for completing all recommended homework.

Although the reports referenced above provide some evidence of the impact of test preparation on outcomes, there are important limitations in the literature. These limitations include the quality of study designs, possible heterogeneity in coaching and coaching impacts, and changes in the tests themselves (Briggs, 2009). While many test-prep studies have focused predominantly on paid coaching and private tutoring (e.g., Briggs, 2009; Appelrouth et al., 2018), the increase of free online test preparation requires

greater attention as online features such as adaptive learning environments and scaled instructional content have the potential to democratize access to an array of educational opportunities, including test preparation (e.g., Means et al., 2010). Finally, there is a paucity of research on the impact of preparation on SAT scores for the revised SAT that launched in 2016, with one analysis demonstrating score gains from PSAT/NMSQT to SAT associated with the use of OSP (College Board, 2018b). The current analysis aims to address important gaps in the literature by providing a detailed description of the practice behaviors of students using OSP and the association of specific practice behaviors with outcomes on the revised SAT.

Official SAT Practice: Product Features and Functionality

In this section, we explain how the skill levels of students are initialized in order to personalize their practice on OSP. The OSP features and functionality described are based on what was available at the time of this study. The numbers in the annotated screengrab of OSP below (see Figure 1) correspond to the subsections below where we describe the core features and functionality of OSP.

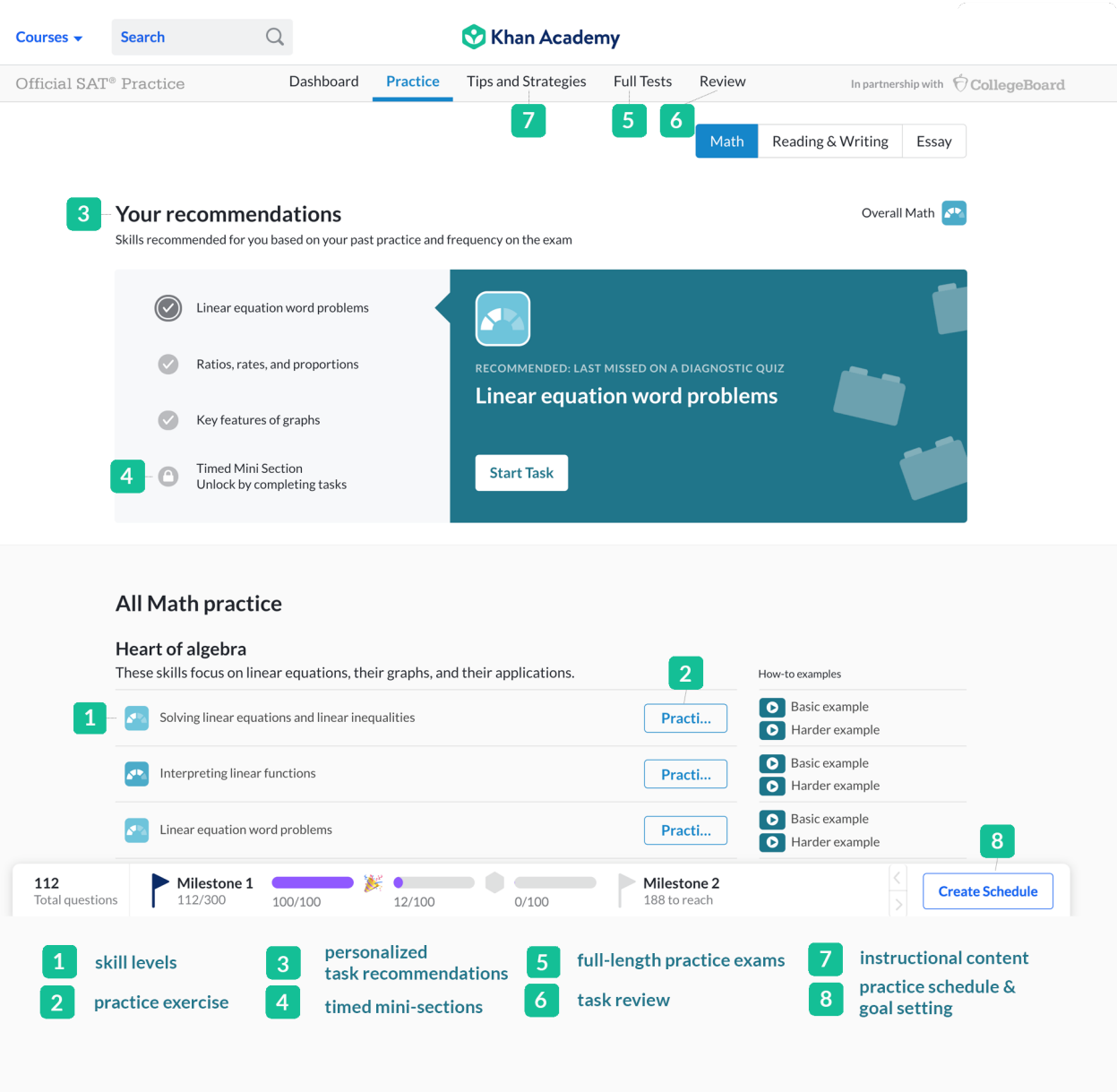


Figure 1. Overview of Khan Academy Official SAT Practice features and functionality (number annotations correspond to product features).

1. **Skill levels.** Content skills and skill levels are a central part of the OSP experience. The content areas assessed on the SAT are discretized into 69 skills on OSP Khan Academy Official SAT Practice: 41 math skills, 7 passage-based reading and writing skills, and 21 grammar skills. For each skill, there are four difficulty levels. Level 1 is foundational (below SAT-level difficulty), and levels 2, 3, and 4 correspond to easy, medium, and hard levels on the SAT, respectively. Skill levels are fundamental to OSP because a student's current level of each skill determines the difficulty of the content that is presented to them on practice exercises. Skill levels are initialized in one of three ways: (1) complete a series of 8 short 10-question diagnostic quizzes on Official SAT Practice (four in math and four in reading and writing). Each quiz assesses a subset of skills in the corresponding domain; (2) complete a full-length practice exam on OSP—item-level

performance for each tested skill is used to determine the student's current level on that skill; and (3) students can link their Khan Academy and College Board accounts. When students link these accounts, they consent to data sharing between Khan Academy and College Board, which enables Khan Academy to access the students' latest SAT Suite of Assessments score data (if any). Specifically, Khan Academy imports item-level data, similar to what College Board reports to students on the Question-Level Feedback portion of their SAT Suite score summary report. The Question-Level Feedback report indicates the difficulty level of each item (easy, medium, or hard) and the student's performance (correct or incorrect). Khan Academy imports the item-level data, coupled with information about the content skill assessed by each item, and uses this data to initialize (or update) the student's corresponding skill levels on Official SAT Practice. To protect student privacy, Khan Academy does not save or store students' SAT Suite data; the Khan Academy API that imports College Board data automatically deletes the imported data once the student's skill levels are set.

2. **Practice exercises.** The practice library contains practice exercises for each of the skills tested on the SAT. Students have the option to direct their study by choosing which skills to practice. Within each practice exercise, if students get stuck or need a refresher, they can access step-by-step hints for each math question (at this time ERW questions do not contain hints, but they will in the future), as well as instructional videos that include worked examples. When a practice exercise is completed, the student's performance on that exercise is used to adjust their level on the corresponding skill.

3. **Personalized task recommendation.** Although students have the option to practice any skill in the practice library, the OSP recommendation engine is designed to help them focus on the skills they would benefit most from practicing. Specifically, the recommendation engine is designed to maximize utility by identifying and prioritizing the skills that a user currently has a low performance on and that are the most likely to occur on the SAT. The recommendation queue for math and for reading and writing always contains four items: three practice exercises and a timed mini-section (explained in the next section). Once those four tasks are completed, the queue is reset with four new tasks: three new practice exercises and a new timed mini-section.

4. **Timed mini-sections.** As the name suggests, timed mini-sections are short mixed-skill practice tasks with a time limit. Timed mini-sections are meant to help students build stamina and practice time management and pacing skills. Timed mini-sections are always part of the recommendation queue but they are initially "locked." In order to unlock a timed mini-section, the student must first complete the three recommended practice exercises in the recommendation queue.

5. **Full-length practice exam.** There are 8 full-length practice exams available on OSP. Six of these practice tests were previously live operational SAT tests; two are never-before-released tests. Students can complete the full practice exams online and pause between sections as needed. Thus, it is not necessary to complete the full 3-hour practice exam in one sitting. Taking a practice exam in one continuous sitting more closely simulates the real test experience. However, the ability to pause between sections makes the full functionality of the practice exams accessible to students who want to take them but who might not have three consecutive hours of access to OSP.

6. **Task review.** Students can review their previously completed practice exercises, timed mini-sections, and full-length exam answers and scores. Once a previous task is selected a student can view all of the problems along with rationales for each answer choice. Math problems also include a fully worked solution to the problem.

7. **Instructional content.** In addition to the in-depth solution steps, hints, and explanations for each question, OSP features worked example videos and lessons and strategy articles. The videos are short, narrated segments showing worked examples of SAT-type problems. Each math and grammar skill in the OSP library has two associated "worked example" videos—one focusing on a more basic problem and one focusing on a more advanced problem. The passage-based ELA skills also have worked example videos. The "Tips and Strategies" tab features a range of articles and videos covering topics such as SAT format, content, question types, and scoring, as well as strategies for time management, active reading, and how to avoid careless errors.

8. **Practice schedule and goal setting.** Students can optionally create a practice schedule based on when they are planning to take the SAT, how many full-length practice tests they want to take before the SAT, and when/how long they want to practice (i.e., which days and how long per day). Khan Academy sends email reminders to students (if they opt in) about their practice schedule, and the schedule is available on their Official SAT Practice dashboard.

Defining 'Best Practice Behaviors' on OSP

As outlined above, OSP offers a rich array of features and functionality to help students prepare for the SAT. Given this functionality and the varied needs of individual students, there is no single "best" or ideal way to use OSP. However, we believe there are several OSP behaviors that are broadly applicable based on how the product is designed, the data elements available, and prior research concerning the effectiveness of test preparation strategies. These behaviors constitute a working operational definition of best practices, not a comprehensive definition. There are several notable behaviors that are not included, such as whether a student set and followed a practice schedule and whether a student spent time reviewing their incorrect responses. These omissions are due to logistical issues; specifically, due to the nature and granularity of data instrumentation for OSP, some usage behaviors are challenging to reliably isolate and quantify. In this study, we operationalize three best practice behaviors on OSP including:

- **Leveling up skills.** Students with linked accounts begin their path through OSP content with a skill level initialized from their previous PSAT/NMSQT performance, as such students are placed at their learning edge when they begin practice on OSP. As students progress through OSP material, they can achieve new levels in the skills practiced, up to a level 4. There are 69 different skills, and not all students will level up on skills in the same way. For example, students may practice very broadly across many skills, without spending enough time on one skill to "level up." However, overall leveling up in at least some skills provides a general signal that students are consistently advancing in a domain tested on the SAT and is a marker for learning progress on OSP**. We operationalize this best practice as students leveling up 15 or more skills (out of 69 total skills) on OSP.**

- **Following skill recommendations.** The OSP task recommender aims to identify the skills a student would benefit the most from practicing by considering the student's past performance and the frequency with which each skill occurs on the SAT. While there are plenty of reasons why a student might also need or want to practice specific skills that are not in their personalized recommendation queue, progressing in recommended skills is a best practice because it guides a student toward efficient use of their OSP time by focusing on skills that are highly relevant to SAT questions and are in the greatest need of attention for an individual learner. **We operationalize this best practice as students completing 10 (or more) practice tasks with the majority of tasks recommended to them.**

- **Completing a full-length practice exam.** Previous research has repeatedly shown that practice tests are an effective strategy for improving test performance and are often more effective than other non-testing learning conditions, such as restudying or exclusive practice (see Adesope et al., 2017 for a recent meta-analysis). The effectiveness of practice tests lies in a combination of cognitive, metacognitive, and noncognitive benefits that occur when simulating the real test, as discussed above. **We operationalize this best practice as students completing at least one full-length practice exam on OSP.**

Methods and Results

Sample

Participants in this study were students in the 2019 U.S. high school graduating cohort who met the following three criteria:

1. Took the PSAT/NMSQT(National Merit Scholarship Qualifying Test) in October of their junior year
2. Took a subsequent SAT prior to graduating (either during their junior or senior year)
3. Linked their Khan Academy and College Board accounts

The first two criteria enable an analysis of SAT performance while adjusting for prior achievement. Students in the 2019 graduating cohort took the 11th-grade PSAT/NMSQT during October 2017. The majority (71%) of these students then took the SAT for the first time the following spring, during the March, April, May, or June test administration dates. On average, the duration of the interval between the students' 11th-grade PSAT/NMSQT and their first SAT was about 26 weeks.

The third criterion functions as a consenting mechanism. When using OSP, a student has the option to link their Khan Academy account to their College Board account. As part of the account-linking process, a student is asked whether they consent to Khan Academy sharing their usage data with College Board. The main user-facing benefit of linking accounts is that doing so is an efficient way to personalize practice on OSP. When a student links accounts, Khan Academy Official SAT Practice is able to access the student's latest PSAT™ or SAT data from College Board. Item-level PSAT and SAT data are then used to make personalized practice recommendations and to present content at the appropriate difficulty level for each skill. Deciding not to link accounts is nonpunitive. If a student chooses not to link accounts, the same degree of personalization within OSP is possible once the student initializes skill levels through one of the alternate means outlined above (e.g., completing diagnostic quizzes or a full-length practice exam on OSP).

Table 1 shows descriptive statistics of the analytic sample for this study, compared to the full population of SAT test takers from the 2019 U.S. high school graduating cohort. About 2.2 million students in the 2019 graduating cohort took the SAT at least once, and 1.3 million took the PSAT/NMSQT in October of their junior year, prior to taking the SAT. Of those 1.3 million students, 545,640 linked their College Board and Khan Academy accounts. Therefore, the linked sample reflects 25% of the full population of SAT test takers and 42% of the population of students who could have used OSP with personalized practice based on their PSAT/NMSQT data. The linked sample is very similar to the full population of PSAT/NMSQT+SAT test takers in terms of race/ethnicity, parental education, and PSAT/NMSQT performance (with the exception that the linked sample has a higher percentage of females and lower percentage of 11th-grade students scoring in the first quartile of the PSAT/NMSQT). All following analyses were conducted with the analytic sample.

Table 1. Demographics of SAT Test Takers from 2019 High School Graduating Cohort

	Subgroup	SAT Test Takers	SAT test takers who took 11th-grade PSAT/NMSQT	Analytic Sample: SAT test takers who took 11th-grade PSAT/NMSQT and linked Khan Academy and College Board accounts
Total		**2,220,087**	**1,291,916**	**545,640**
Gender	Female	52%	53%	58%
	Male	48%	47%	42%
Race/ Ethnicity	American Indian / Alaska Native	1%	<1%	<1%
	Asian	10%	10%	11%
	Black / African American	12%	11%	11%
	Hispanic / Latino	25%	27%	27%
	Native Hawaiian / Other Pacific Islander	<1%	<1%	<1%
	White	43%	46%	44%
	Two or More Races	4%	4%	4%
	No Response	5%	2%	2%
Parental Education	No High School Diploma	9%	9%	9%
	High School Diploma	27%	27%	27%
	Associate Degree	7%	7%	7%
	Bachelor's Degree	28%	31%	31%
	Graduate Degree	21%	24%	24%
	No Response	8%	3%	2%
11th-Grade PSAT Quartile	Q1 [320–910]	---	29%	24%
	Q2 [920–1050]	---	25%	26%
	Q3 [1060–1180]	---	22%	25%
	Q4 [1190–1520]	---	24%	26%

1. Descriptive Findings of OSP Usage Patterns

Key Findings

Roughly 10% of students spend 6+ hours or complete a best practice on OSP.

Students use OSP primarily within the two months before they take the SAT.

Best practice behaviors are correlated, but students show selective strategies for how they engage with OSP.

In this section, we explore the overall usage patterns observed on OSP to understand how students utilized OSP features and to obtain insight into potential group differences in OSP usage. For definitions of variables, see Appendix A. We explore the depth of overall usage by examining the frequency with which students use different features of the tool, how students' usage differs across different best practices, and if certain subgroups of students are more likely to engage in best practices.

1a. What does student usage on OSP look like?

Roughly 10% of students spend six or more hours or complete a best practice on OSP.

Figure 2, the Depth of Usage chart, shows the percentage of students engaging in various best practice behaviors on OSP along with the percentage spending significant time on the platform. We show this usage both for all linkers, as well as for linkers who go through at least one problem on OSP, as a significant subgroup of students take the initial step of linking their accounts but never complete any time on OSP.

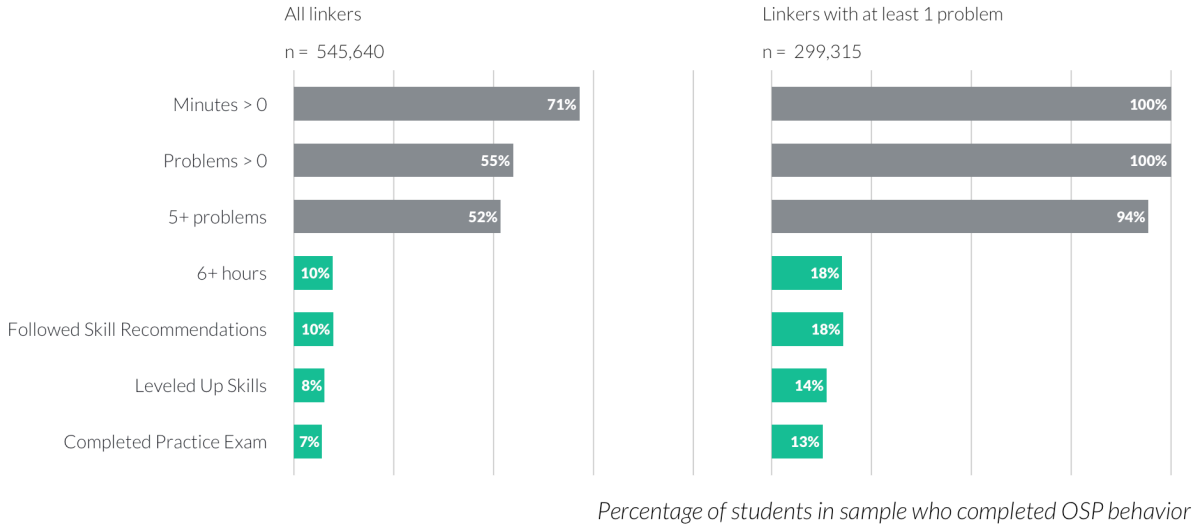


Figure 2. OSP depth of usage.

Table 2 shows an overall summary of how students interacted with OSP, broken down by demographic groups and PSAT performance groups. This table provides a reference for how the overall usage varies across subgroups within our sample. We provide this table as a descriptive reference; in section 2b we present analyses to determine if there are meaningful differences in usage across subgroups.

Table 2.
OSP Usage by Subgroups for Linkers who Completed at least One Problem.

					Target OSP usage behavior			
group	**subgroup**	**n**	**prct**	**Median hours**	**6+ hours**	**Leveled Up Skills**	**Completed Practice Exam**	**Followed Skill Recommendations**
Total (Linkers with at least 1 problem)		299,315	100%	1.8	18%	14%	13%	18%
Gender	Male	121,910	41%	1.9	18%	14%	13%	18%
	Female	177,405	59%	1.8	17%	14%	13%	18%
Race/Ethnicity	White	132,388	44%	1.7	16%	15%	14%	19%
	American Indian or Alaska Native	1,100	0%	1.5	13%	6%	10%	14%
	Asian	33,722	11%	2.7	28%	20%	17%	24%
	Black or African American	34,383	11%	2	20%	10%	11%	15%
	Hispanic or Latino	78,724	26%	1.7	16%	10%	10%	15%
	Native Hawaiian or Other Pacific Islander	524	0%	1.5	15%	8%	9%	17%
	No response	4,939	2%	2	21%	15%	13%	21%
	Two or more races	13,535	5%	1.8	18%	15%	14%	20%
Parental education	No High School Diploma	25,451	9%	1.7	16%	9%	10%	14%
	High School Diploma	40,963	14%	1.6	15%	10%	10%	15%
	Associate Degree	60,856	20%	1.7	15%	11%	11%	15%
	Bachelor's Degree	93,576	31%	1.8	18%	15%	14%	19%
	Graduate Degree	72,316	24%	2.1	22%	19%	16%	22%
	No response	6,153	2%	1.6	16%	10%	10%	16%
PSAT quartile	Quartile 1 [320, 920)	66,641	22%	1.6	14%	5%	8%	13%
	Quartile 2 [920,1060)	76,932	26%	1.6	15%	9%	10%	15%
	Quartile 3 [1060,1190)	75,069	25%	1.8	18%	15%	13%	18%
	Quartile 4 [1190,1520]	80,673	27%	2.2	23%	24%	19%	25%

Commensurate with the low rates of completion and high rates of dropoff that research has documented across many free online learning platforms with large usage (Gütl, Rizzardini, Chang & Morales, 2014; Kizilcec & Halawa, 2015), a relatively small percentage of students in our sample spent six or more hours

on OSP (10% of all linkers). However, for students who engaged with the platform beyond linking and completed at least one problem, this six or more hours usage group rises to 18%. Considering best practice behaviors other than time spent on OSP, we see that as the time required to complete them increases, smaller percentages of students engage in these more time-consuming behaviors (e.g., completing a full-length practice exam vs. completing 10 recommended tasks).

1b. When are students using OSP?

Primarily within the two months before they take the SAT.

As testing dates and OSP usage are subject to student and school choices, there are variable amounts of time between the students' PSAT/NMSQT and first SAT test dates. One concern when evaluating OSP usage as a contributor to SAT performance is that it is possible that some students' time on OSP occurs far in advance of their SAT testing date. In this case, we might question whether OSP truly serves as an effective intervention for SAT performance.

Figure 3 shows the percentage of students' minutes on OSP, and how those percentages are allocated in the lead-up to the SAT. Crucially, this figure illustrates that most student activity on OSP is concentrated in the weeks immediately leading up to the SAT.[1] Among students who used OSP for six or more hours, the average student spent 80% of their total OSP time within 8 weeks of the SAT and nearly all (98%) of their time within 12 weeks.

[1] It is still important to note the variance in students' time between PSAT/NMSQT and SAT across this sample, which may drive differences in students' performance between their PSAT/NMSQT and SAT scores (e.g., students will necessarily have different amounts of time to study between tests). We consider this further by including the number of weeks between test dates as a covariate in our predictive models.

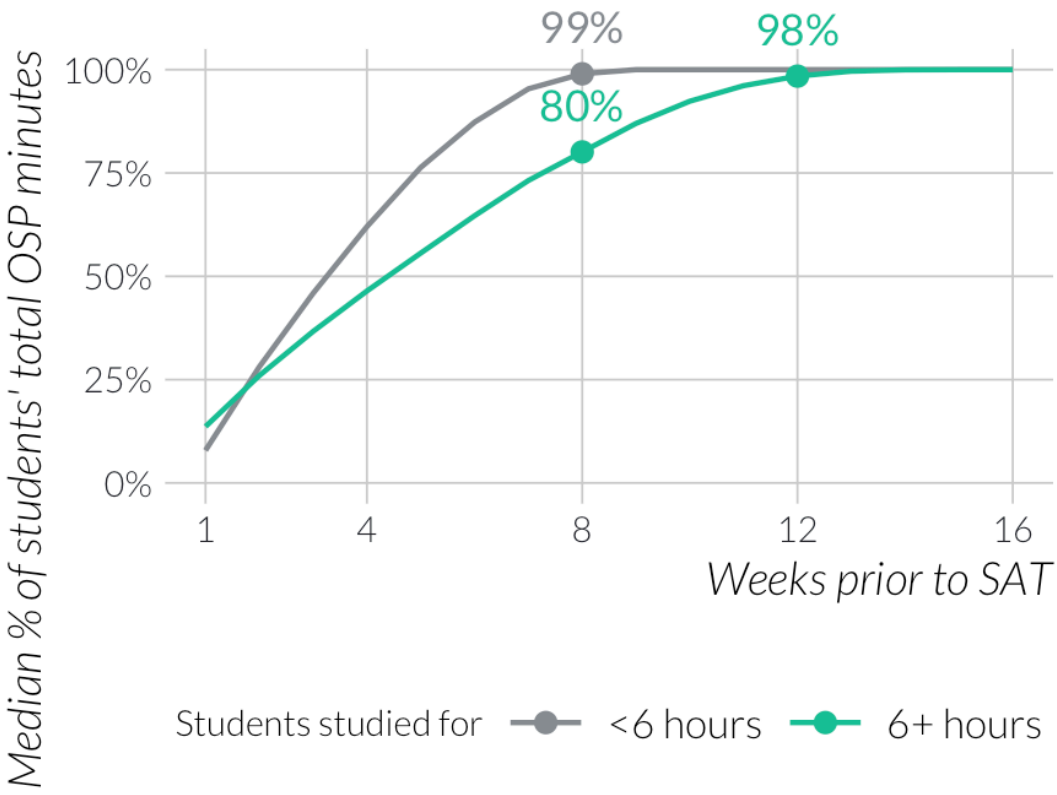


Figure 3. Concentration of OSP practice minutes between PSAT/NMSQT and SAT, among linkers who attempted at least 1 problem.

1c. How are students engaging in best practices?

Best practice behaviors are correlated, but students show selective strategies for how they engage with OSP.

In this subsection, we explore how students engaged with the above described best practice behaviors within overall usage. Figure 4 shows the strength of co-occurrence between any two of the OSP usage variables for linkers who completed at least one problem on OSP. There is a positive correlation between spending six or more hours and each of the other best practice behaviors (completing a full-length practice exam; consistently following practice task recommendations; and leveling up).[2] On the surface, this is not surprising since those practice behaviors require time. Notably, however, the magnitude of these correlations is moderate, indicating that some students who spend six or more hours do not do any

[2] In order to show the best practice behaviors as defined above, Figure 4 presents the bivariate correlations between several dichotomized variables: i.e., the best practice measures are calculated from the raw platform usage data and assigned as categorical measures to students. It is possible that correlating dichotomized variables may dampen the strength of the relationship between measures. However, in this case, when we examined correlations between the raw variables which make up the best practice behaviors we found closely similar results.

of the best practice behaviors and/or that some students do the best practice behaviors without spending six or more hours.

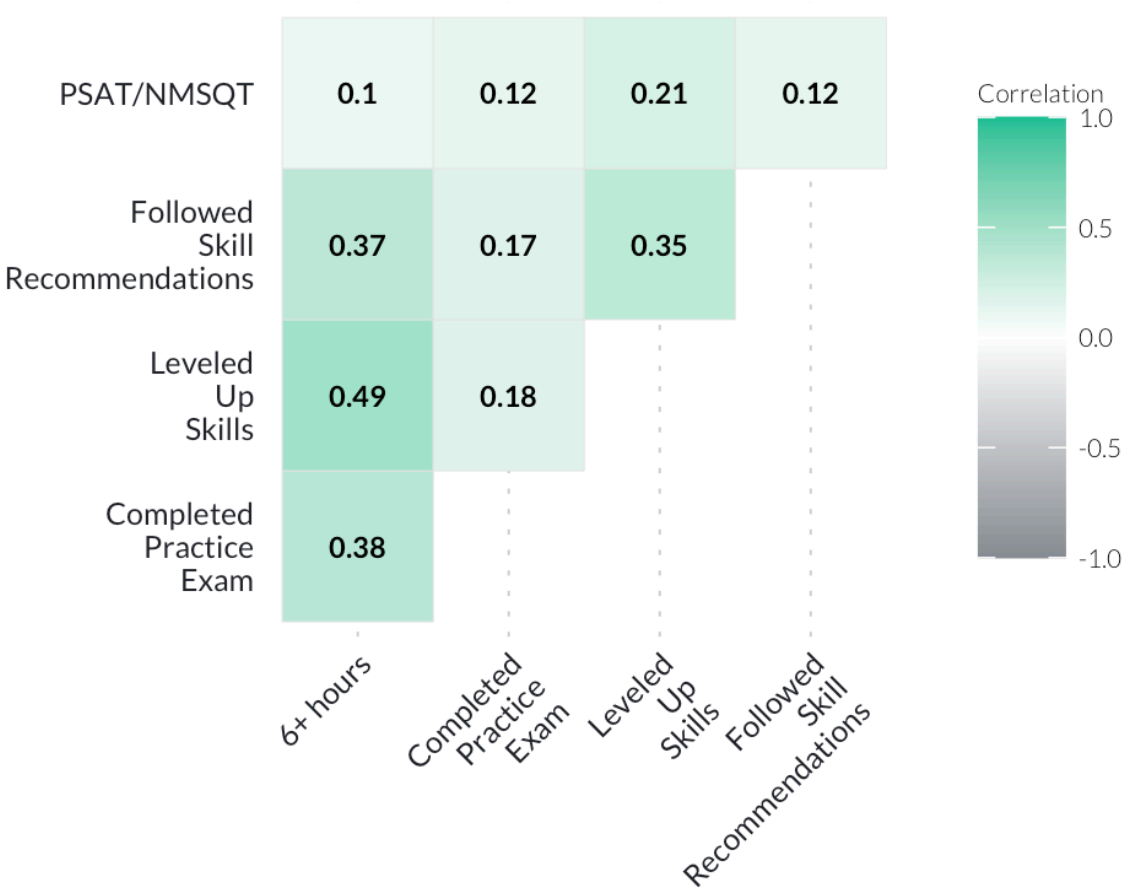


Figure 4. Bivariate correlation between OSP usage variables for linkers who completed at least 1 problem.

Next, we examined the frequency of all possible combinations of the best practice behaviors occurring. As OSP provides a wide range of possible uses, students with different needs may choose to emphasize different features. Thus, it is important to understand the different ways students engage in best practice behaviors.

Best practice behaviors are not expected to be distinct and independent user patterns on OSP, as many of these behaviors overlap, and working toward one best practice may impact another. Increased time spent on the platform is also necessarily associated with certain best practices: for example, completing a full-length practice exam requires students to spend three hours on OSP, and leveling up in initial skills requires a minimum amount of time spent practicing those skills. However, students who engaged in one best practice behavior would not necessarily engage in the others equally, nor would best practice behaviors increase equally with increased OSP usage.

Figure 5 provides a more granular view of the frequency with which each behavior co-occurs with the others across the entire sample. In this figure, we divide usage into several groups that will be repeated in section 2c as an exploration of the best practices.

The vertical bars in this figure map to our overall sample and display a funnel of usage, with dark grey representing students with no OSP usage and lighter grey representing students who spent less than six hours on OSP and did not complete at least one best practice behavior. The colorful bars show students who engaged more with OSP: the red bar represents learners who spent six or more hours on OSP but who completed no best practices, gold bars are learners who similarly spent six or more hours but did show at least one best practice, and blue bars are learners who showed at least one best practice but within an overall usage of less than six hours.

Within these usage groups, we can also see the frequency of all possible combinations of behaviors, and the connected closed circles indicate the combinations of the best practice behaviors co-occurring. The horizontal bars summarize the frequency of each best practice behavior in the overall data. This intersection plot shows the relationships between the best practice behaviors that we have operationalized as useful signals in this study, mapping the frequency of that behavior along with its co-occurrence. This shows the number of students who demonstrated given combinations of best practice behaviors.

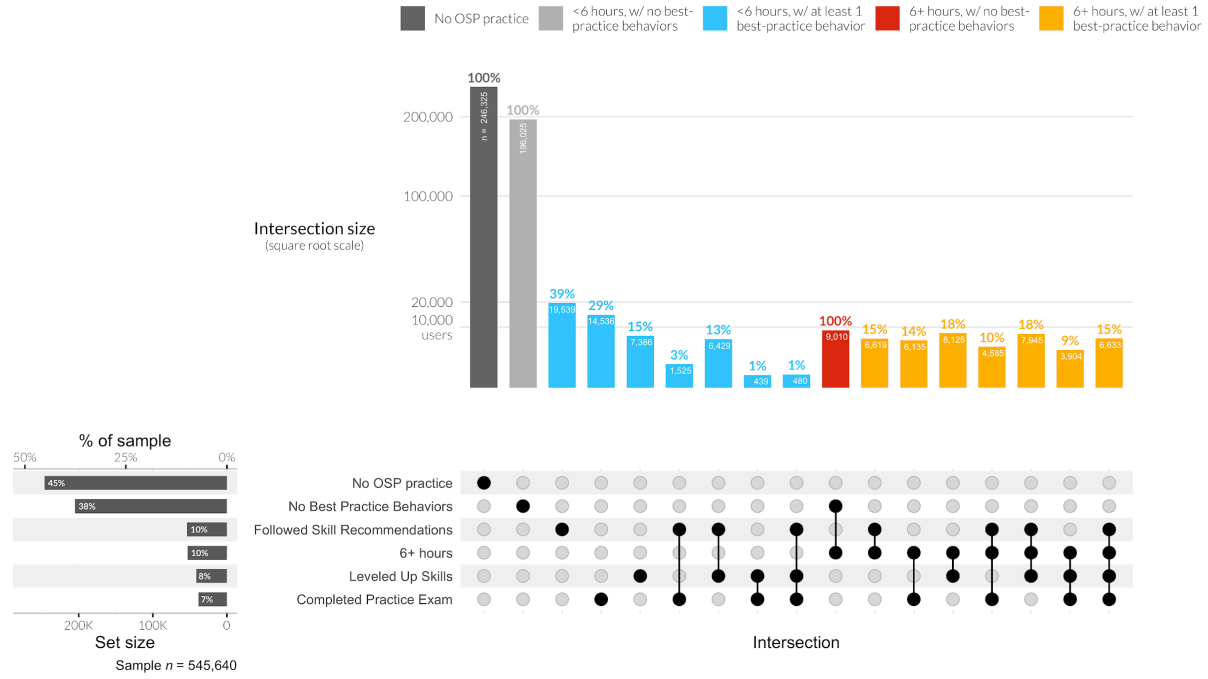


Figure 5. Intersection of best practice behaviors for linkers.

A few relevant patterns emerge from these descriptives: a sizable subgroup of students completed at least one practice exam, but did not spend six or more hours on OSP, suggesting that they were primarily using the tool to access practice exams. Another sizable subgroup spent six or more hours on OSP without engaging in any of the other best practice behaviors, suggesting that they did not focus on recommended tasks, complete practice exams, or level up skills. Another group of students leveled up in more than 15 skills but did not spend six or more hours on OSP, suggesting that they may be leveling up quickly in many skills but not pursuing the most challenging content for these students.

Overall, best practices are positively associated with each other, as expected (see Figure 4). Yet, there was some suggestive evidence that students could fall into distinct groups that focus on different best practices. It is likely that these different usage patterns emerge because students prioritize OSP features differently depending on their needs and goals. It is possible, for instance, that some students were

encouraged or even required to complete a practice exam on OSP, while other students may have had access to practice exams outside of the platform. Without a deeper understanding of learner contexts, it is out of the scope of this report to fully investigate what determined this pattern of OSP usage. But this pattern suggests that many students will engage differently with a feature-rich platform, and that explicitly encouraging best practice behaviors may be a useful strategy to help learners.

2. Associations Between OSP and SAT Performance

In this section, we broadly examine how use of OSP is associated with SAT performance. We explore this association by examining the relationships between SAT scores and the amount of time that students spent on OSP, as well as the types of behaviors performed by students on OSP during that time. We further break down the relative contribution of best practice behaviors. Throughout, these analyses include important student characteristics such as parental education and gender, along with administrative characteristics of the tests themselves (e.g., the amount of time between a student's PSAT/NMSQT and SAT). Finally, this section explores how student characteristics interact with OSP, allowing us to examine if these interactions work together to strengthen or weaken the relationship to SAT performance.

Key Findings

Spending time on OSP is associated with greater scores on the SAT; 6 hours is associated with an additional 21 points (.11 effect size) more than students who did not use OSP. These findings hold true regardless of student demographic characteristics.

How students spent their time on OSP matters: Students who used OSP for 6+ hours *and* demonstrated at least 1 best practice behavior scored an additional 39 points (.20 effect size) more than students who did not use OSP. This holds true regardless of student demographic characteristics.

Not all subgroups of students are as likely to use best practice behaviors on OSP.

As we examine questions in this section, we present observational analyses of the data designed to leverage the natural fluctuations in student usage of OSP to make inferences about the effectiveness of OSP. We implement various statistical controls to account for the influence of confounding factors. While a true experiment with random assignment to conditions would provide the highest standard of evidence, it was not practically possible to implement such a design in this setting. Nevertheless, this analysis is an important first step in establishing the effectiveness of OSP.

2a. Does time spent using OSP relate to SAT achievement?

Time spent using OSP was associated with positive improvements to SAT performance.

We hypothesize that use of OSP improves SAT performance, as such it is reasonable to predict that there should be a positive relationship between the amount of time that students use OSP and their SAT scores. At a high level, we examined this relationship using a series of multiple linear regression models. We examined composite SAT scores, as well as the math and ERW (evidence-based reading and writing) subscores. The goal of this analysis was to estimate the difference in SAT score for students who spent more time using OSP while controlling for as many confounding factors as possible. The most important factor we accounted for was PSAT/NMSQT performance, which represents students' prior achievement before their practice on OSP. We used the composite, math, and ERW PSAT/NMSQT scores when

predicting the composite, math, and ERW SAT scores, respectively. We controlled for several demographic factors; specifically gender, race/ethnicity, and highest level of parental education (e.g., high school diploma, some college, etc.). We also controlled for test-taking conditions, such as whether students took the exam during a school day administration or weekend, and the time interval between the PSAT/NMSQT and the SAT.

For the purposes of this report, we focus our discussion on the estimated impact of time using OSP on SAT performance. For a full breakdown of our statistical modeling procedures and complete results from our regression analysis, see Appendix B. The model estimates for hours spent using OSP are represented graphically on Figure 6. This figure shows the estimated improvement to SAT performance for composite (Panel A), math (Panel B), and ERW (Panel C) scales, as a function of the number of hours spent using OSP, controlling for the confounding variables in the model. In general, we see that the amount of time spent using OSP was positively associated with higher SAT performance. This positive relationship was found at the composite level, as well as each of the math and ERW subscales. However, we do note that the overall impact of OSP usage was slightly larger for math than for ERW, which is consistent with past studies evaluating the effects of test prep (Briggs, 2009). Moreover, the benefits of OSP use taper off more quickly for ERW than math, suggesting that the ceiling of benefits from using OSP is reached more quickly for ERW than for math. It is not clear whether this represents a limit of the OSP platform, or a general limit on the ability to improve ERW performance. It is also worth noting that the top of each panel in Figure 6 represents the frequencies with which students use OSP at given time intervals. The vast majority of students (approximately 80%) use OSP for less than 3 hours. Thus, while increased usage of OSP was associated with increased SAT performance, the majority of students are not using OSP enough to obtain meaningful benefits to their performance.

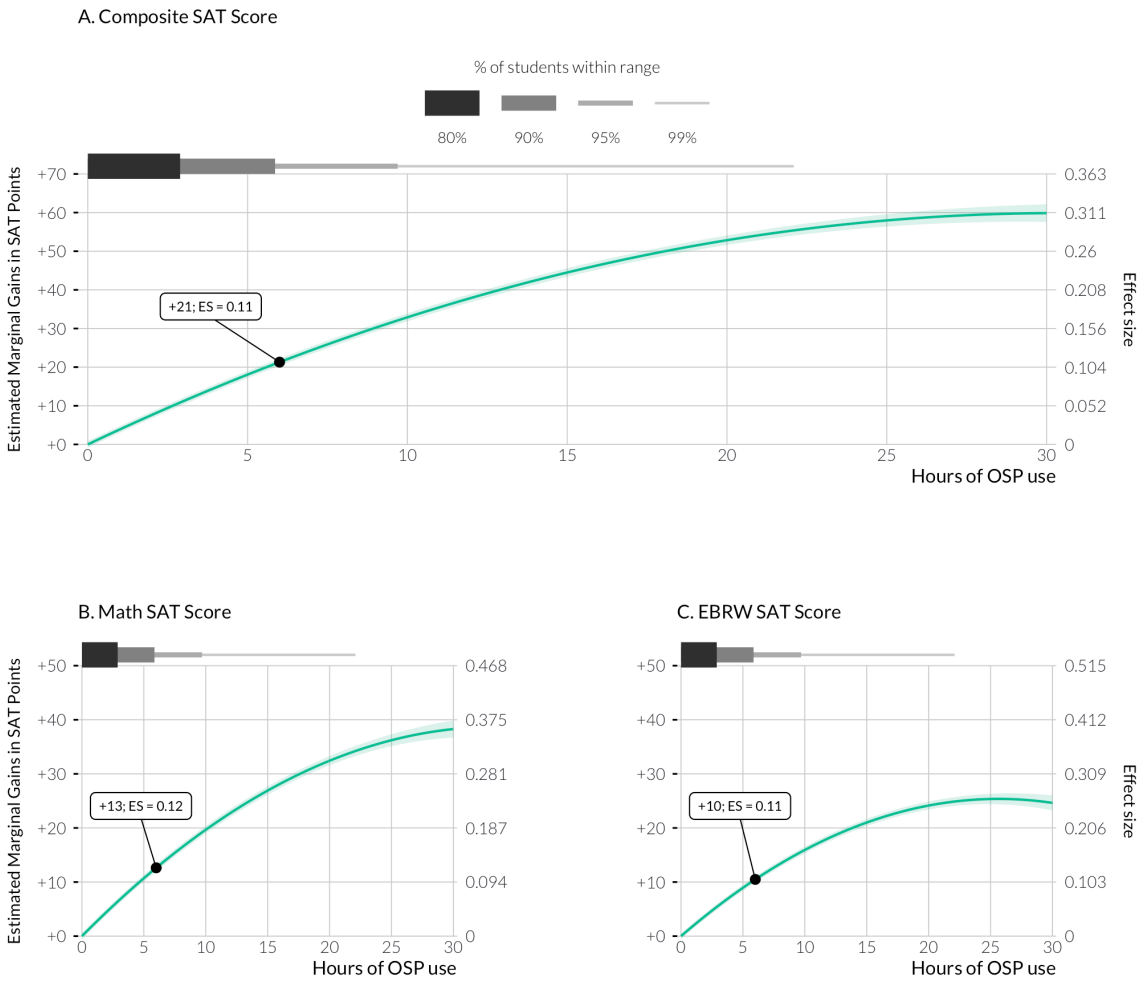


Figure 6. The estimated change in SAT scores as a function of hours using OSP, after controlling for students' PSAT score and demographic characteristics. The effects on composite, math, and evidence based reading and writing (ERW) are shown in panels A, B, and C, respectively. The plots show the increase in SAT points achieved, relative to students who use OSP for 0 hours. The effect size is the change in SAT divided by the overall standard deviation.

In order to estimate the magnitude of the effect of OSP usage, effect sizes are included on the right side of the Y axis in each panel in Figure 6. The effect size is the estimated change in SAT points divided by standard deviation of SAT. In the context of educational interventions, less than .05 is considered small, .05 to .20 is considered medium, and greater than .20 is considered large (Kraft, 2018). For the composite, math, and ERW scales, a large effect size of .20 was achieved at 12.3, 11.1, and 13.3 hours of OSP usage, respectively. At the six hour point, the effect sizes were for 0.11, 0.12, and 0.11 for composite, math, and ERW scales, respectively. These results suggest that OSP usage has slightly greater benefits for math than for ERW, which is consistent with past research on the benefits of test prep. It is important to note that limitations in our data set prevent us from making strong claims on this matter. In order to properly evaluate the differential effects of usage on the SAT subscales, we would need to be able to match the topic domain of the OSP practice data to that of the outcome. Unfortunately, our OSP usage data does not specify the domain in which a student was using OSP, though we do plan on conducting such analysis

when the data become available. For this reason, we will only focus on the effects of OSP at the composite level for the remainder of this report.[3]

2b. Do all students benefit equally from their time spent on OSP?

Students who use OSP appear to show positive benefits, regardless of gender, ethnicity, parental education, or PSAT level.

As an organization, the mission of Khan Academy is to provide a free, world-class education to anyone, anywhere. For this reason, it is critically important to evaluate whether all students benefit equally from using OSP. To this end, we conducted a series of analyses designed to examine whether demographic and prior achievement factors interacted with use of OSP. More specifically, we wanted to know whether students from different groups who spent approximately the same amount of time on OSP achieved different outcomes. Similar to the analysis in Section 2a, we estimated the impact of OSP over time, except we explored whether the impact changed as a function of various student characteristics. As in previous analyses, we controlled for the same confounding variables (PSAT, gender, ethnicity, parental education, test day, weeks since PSAT). For the purposes of brevity, we focus only on the key findings here, specifically, the most highly observed subgroups in each category (e.g., Asian, Black, Latinx, and White for race/ethnicity). Estimates for under observed categories had too much uncertainty to draw reasonable conclusions. The full details of our analysis and model results are in Appendix C.

The impact of OSP across levels of gender, ethnicity, and parental education are shown on panels A,B, and C of Figure 7, respectively. In each of the panels, we can observe several performance differences at the group level. These differences are commonly observed in the education literature, and theorized to result from the "opportunity gaps," which are also documented for these groups. Therefore, it is not surprising to replicate this difference in our sample. However, it is still important to examine whether there is evidence that OSP usage is associated with different outcomes when comparing between these groups.

For evaluating the impact of OSP specifically, the critical information is contained in the slope and shape of the curves representing the rate of increase in SAT performance over time. When examining the overall slope of these curves, we see that all subgroups derive a positive benefit from increased usage of OSP regardless of gender, ethnicity, parental education, or PSAT score. Any differences in the benefits are only slight. For example, in Figure 7a, the increasing benefits of usage for females tapers off more rapidly than males. However, it can be difficult to gauge how meaningful this difference actually is. To aid in this regard, we included point estimates of the benefit for each group at the six-hour point. Remembering that six hours is the recommended minimum amount of time to spend using OSP serves as a useful checkpoint for comparing the relative effects observed across the groups. For example, the biggest difference between ethnicity groups at the six hour mark is only an estimated 7 points. Through this lens, any of the observed between group differences are not practically significant.

[3] One possible concern with the composite SAT measure is whether using a composite score will mask divergent subscores in math and ERW: i.e., a student with a high ERW score and a low math score might have the same composite score as a student who scored near the mean for both subsections. However, math and ERW scores in our sample were strongly positively associated (.78), indicating that overall students' subscores were not systematically divergent.

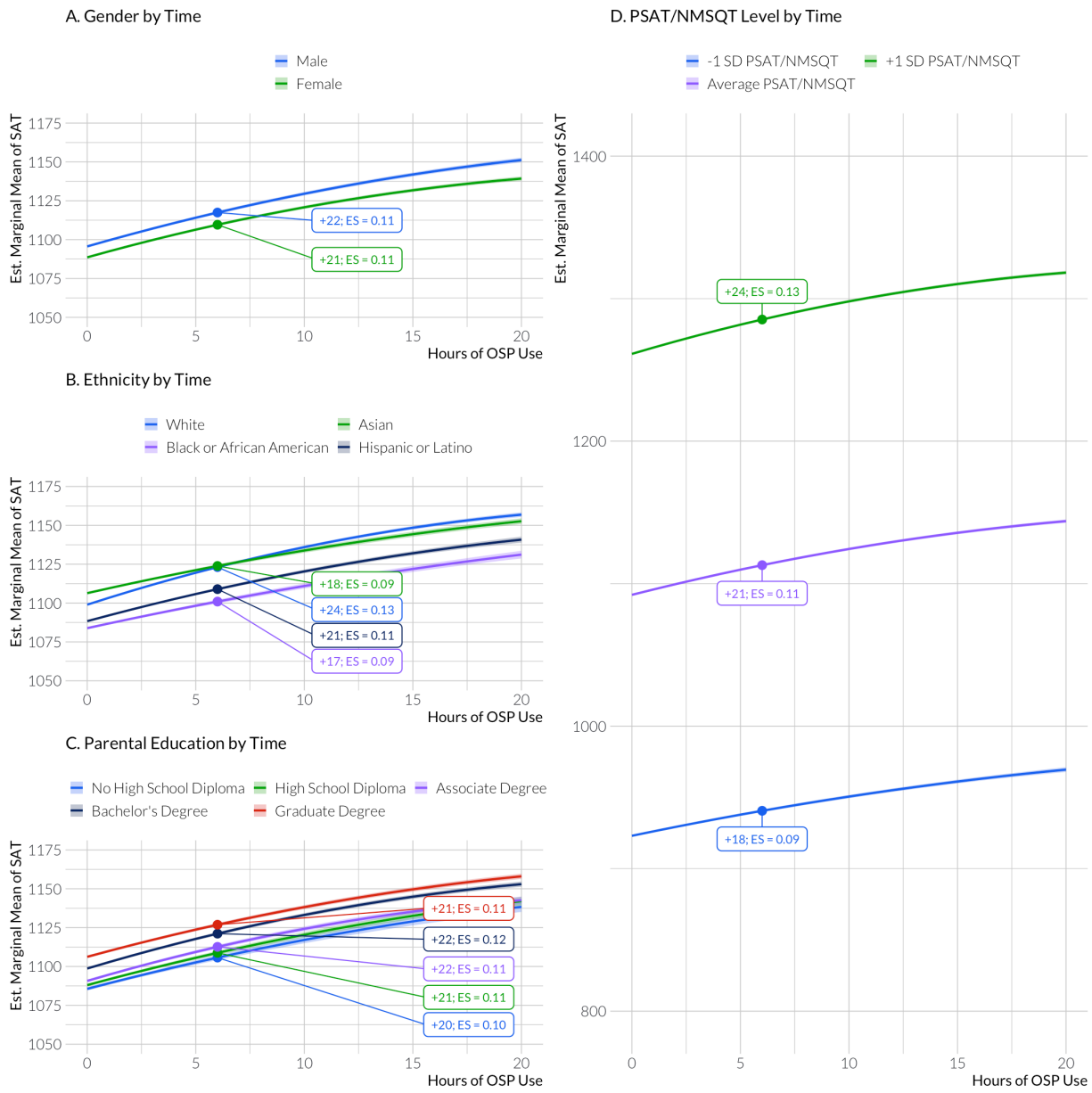


Figure 7. The estimated impact of using OSP as a function of student characteristics; gender (Panel A), ethnicity (Panel B), parental education (Panel C) and PSAT level (Panel D). Estimated marginal means are shown on the Y axes, which correct for other confounding variables. Effect sizes (ES) for each group are shown at the six-hour point. Effect size is the change in SAT from 0 hours of use to 6 hours, divided by the overall standard deviation (see Appendix B for calculation). Shaded regions are 95% confidence intervals.

Next, we consider the impact of OSP across levels of prior achievement, as measured by PSAT/NMSQT scores. Given the wide range of possible PSAT/NMSQT values, plotting this interactive effect is not as straightforward as the demographic variables. To aid in this regard, we plot the effects of OSP usage at three levels of composite PSAT/NMSQT scores; one standard deviation below the mean (872), the mean (1059), and one standard deviation above the mean (1246). The effects across each of these three groups are shown on Figure 7d. Again, we see that all groups showed positive effects of using OSP. There were small differences on the overall slopes for each of these levels. In general, students scoring 1 standard

deviation above the mean PSAT/NMSQT scores showed the steepest increases in SAT achievement, followed by those scoring at average PSAT/NMSQT, than those scoring 1 standard deviation below the mean PSAT/NMSQT. Comparing performance at the six hour mark for reference, the 1 standard deviation above the mean students gained roughly 26 points, whereas students with 1 standard deviation below the mean PSAT/NMSQT scores gained a more modest 18 points. However, these differences are very small, and probably do not reach a threshold of being practically meaningful since the PSAT/NMSQT is scored in increments of 10. The important thing to note is that students at all levels demonstrated an ability to improve their performance by using OSP.

2c. Are best practice behaviors associated with improved SAT performance?

Students who used at least one-best practice behavior outperformed students who spent a similar amount of time on OSP.

In the previous section, we demonstrated that the overall amount of time using OSP was related to higher SAT performance. Of course, time alone does not cause a person to learn. We hypothesize that how people spend their time on OSP has differential associations with learning. To test this, we examine if students who spend their time engaging with best practice behaviors achieve better outcomes on the SAT than students who do not. In the above section, we outlined three specific best practice behaviors: (1) leveling up skills, (2) completing a full-length practice exam, and (3) following skill practice recommendations. In this section, we will examine how these best practice behaviors contribute to overall achievement on the SAT.

To examine the overall impact of the best practice learning behaviors, we used a linear regression model similar to those used in Section 2a. The goal of this analysis was to determine if students with varying hours of usage on OSP and at least one of the three best practice behaviors perform better on the SAT than students who do not. Specifically, we classified students into one of five groups:

1. Linked, but no OSP Practice
2. Less than six hours OSP, no best practice behaviors
3. Less than six hours OSP, at least one best practice behavior
4. Six or more hours of OSP, no best practice behaviors
5. Six or more hours of OSP, at least one best practice behavior

At least one best practice behavior meant that the student met the minimum threshold required for any of the three best practice behaviors. In our analysis, we used the same control variables as those used in the previous analysis (i.e., PSAT/NMSQT score, demographics, test conditions). Note that due to limitations in our ability to tie the specific learning behaviors to a topic domain, we only focused on composite SAT achievement in this analysis.

For the purposes of this report, we focus our discussion on the estimated additional SAT score increase seen in the respective OSP practice groups. For a full breakdown of our statistical modeling procedures and complete results from our regression analysis, see Appendix D. Model estimates for the practice groups, relative to the "No OSP Practice," are shown in Figure 8. The Y axis on the left shows the additional SAT points predicted from the OSP usage groups defined above, relative to the no OSP practice group. Focusing first on students who used no best practice behaviors (gray bars), we see the same relationship between time spent on OSP and achievement that we presented in Section 2a: more time using OSP was associated with higher SAT scores. However, when we also examine students who used at least one best practice behavior (green bars), we see that how time is spent using OSP matters greatly. Students who spent fewer than six hours on OSP but used at least one best practice behavior

achieved roughly the same benefit as students who spent more than six hours but did not use a best practice behavior. Moreover, for students who spent more than six hours on OSP, the benefit to their SAT score of that time spent nearly doubled when they performed at least one best practice behavior. These students were estimated to have gained approximately 39.2 additional SAT points relative to the No OSP Practice group—an effect size of .20. We conducted a follow-up analysis to determine whether this benefit was observed for all students, similar to Analysis 2b. The details of this analysis are in Appendix E. To summarize, we did not observe substantive differences in the effects of meeting the six hour and one best practice behavior for students of various background characteristics.

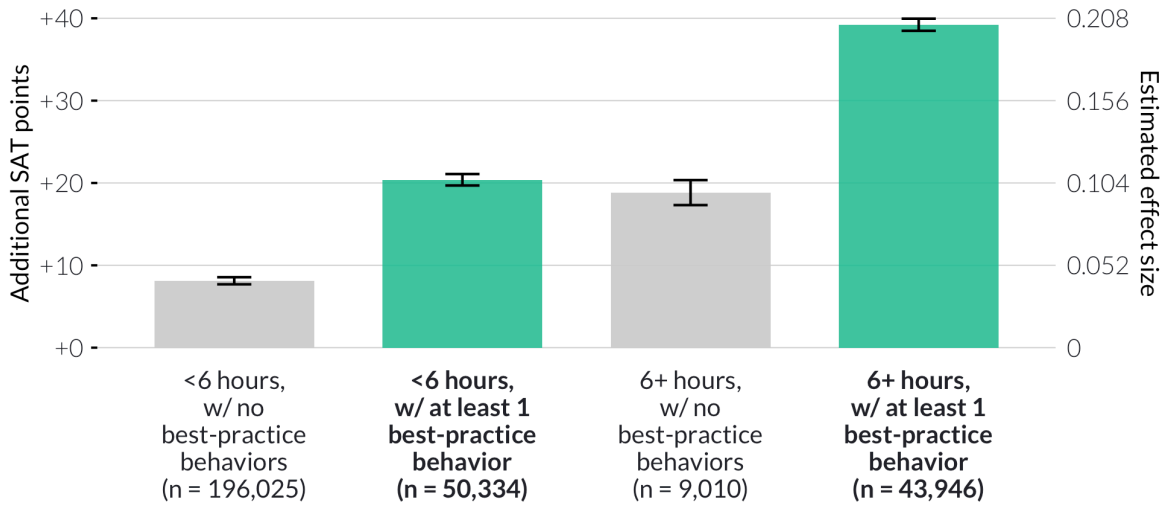


Figure 8. The estimated change in composite SAT scores as a function of OSP usage, after controlling for students' PSAT score and demographic characteristics. The plot shows the increase in SAT points achieved, relative to students who use OSP for 0 hours. Effect size is the change in SAT divided by the overall standard deviation.

It is worth briefly discussing the sample sizes of each of the best practice groups. The reference level "No Practice Group" (not shown in Figure 8), was the largest group in this sample (n = 246,325). For students who used OSP for less than six hours, the majority of students did not engage in one of the best practice behaviors. The situation was reversed for students with more than six hours using OSP, where the majority of students did engage with one of the best practice behaviors. The relationship between time and engaging in a best practice behavior is complicated and was previously discussed in Question 1c.

In this analysis, we attempted to control for the effects of confounding variables using linear regression to fully partial out the effects of these covariates. To scrutinize these results with additional rigor, we repeated this analysis using several propensity score based approaches, which are in Appendix F. The results from these covariate-balanced models are consistent with the effect reported in Figure 8. Thus, we are reasonably confident in the reliability of this estimate.

Relative contribution of the best practice behaviors

While the above analysis makes a strong case that the overall amount of time spent on OSP matters less than the way in which that time is spent, it does not speak to the relative benefits of the specific best practice behaviors. For example, does taking a practice exam produce the same level of benefits as leveling up skills? To this end, we also examined the relative contribution of each by using linear regression to estimate the effects of these behaviors on overall SAT performance, while controlling for

the overall time using OSP and other confounding variables. The full details and results of this analysis are in Appendix D. The key results are shown on Figure 9.

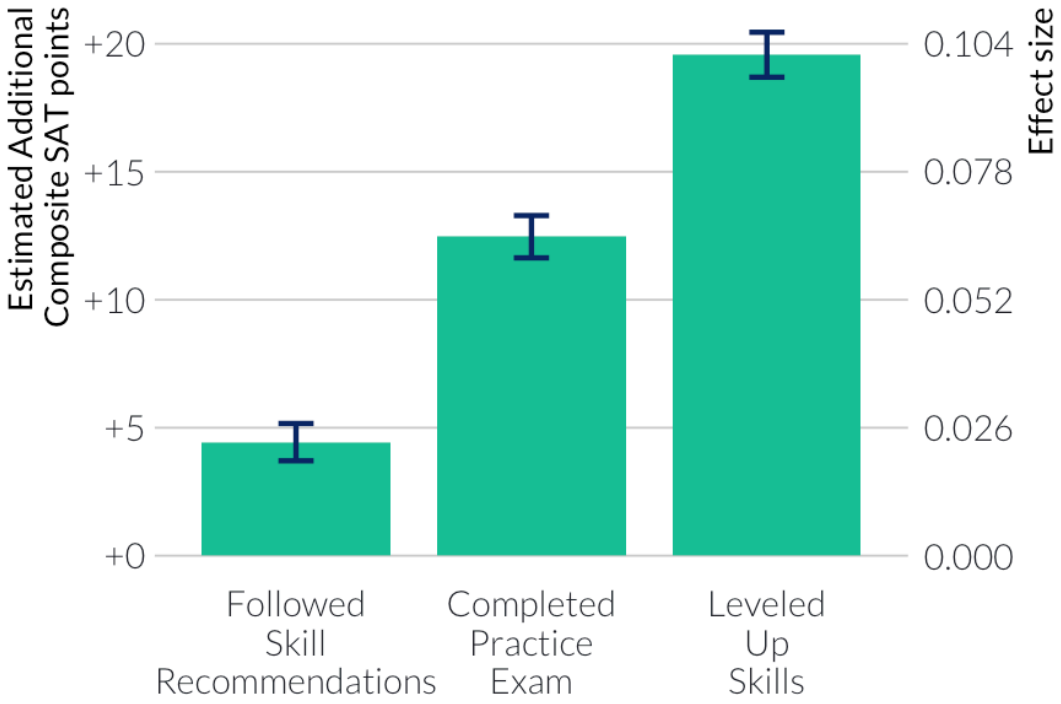


Figure 9. The estimated effect of specific practice behaviors on composite SAT scores, after controlling for confounding variables. The bars show 95% confidence intervals.

Figure 9 shows the estimated marginal effects of each best practice behavior while holding the other behaviors and confounding variables constant. We see in the figure that leveling up skills produced the biggest benefit to SAT performance, an estimated 19.5 additional composite SAT points (ES = 0.10). Comparatively, following practice recommendations tasks resulted in the smallest benefit for SAT performance (+4.56 points; ES = .02). Recall that following practice recommendations meant that students were specifically practicing skills that were recommended to them by OSP, while leveling up skills was more general, in that it applied to any of the skills they leveled up. Thus, the relatively small effect of following recommended practice makes sense given that it is essentially a modification of an already beneficial practice. Moreover, it is important to note that the estimates of these effects take into account the effects of the other variables, and there is certainly a degree of redundancy there. For example, following best practice recommendations may also lead to more leveling up, thus making it difficult to completely tease apart the relative contributions of each. Recall that overall, best practice behaviors are moderately correlated. Nevertheless, the important takeaway is that while more time spent using OSP may be related to SAT performance, the manner in which that time is spent matters considerably. Engaging in practice, via leveling up skills or completing a full-length practice exam, was

associated with the largest effects, whereas following practice recommendations provided small but positive benefits.

Figure 10 illustrates the above points. In Figure 10, the estimated marginal mean score gain across our usage groups are shown. Figure 10 repeats the information found in Figure 5 above, where we outline the relative frequency of best practice behaviors and their combinations within the five usage groups of our sample (including no OSP practice). However, Figure 10 includes the additional SAT points that we estimate for these combinations of both time on the platform and completion of one or more best practices. Visually, this figure shows the increasing benefit of effectively spent time on OSP.

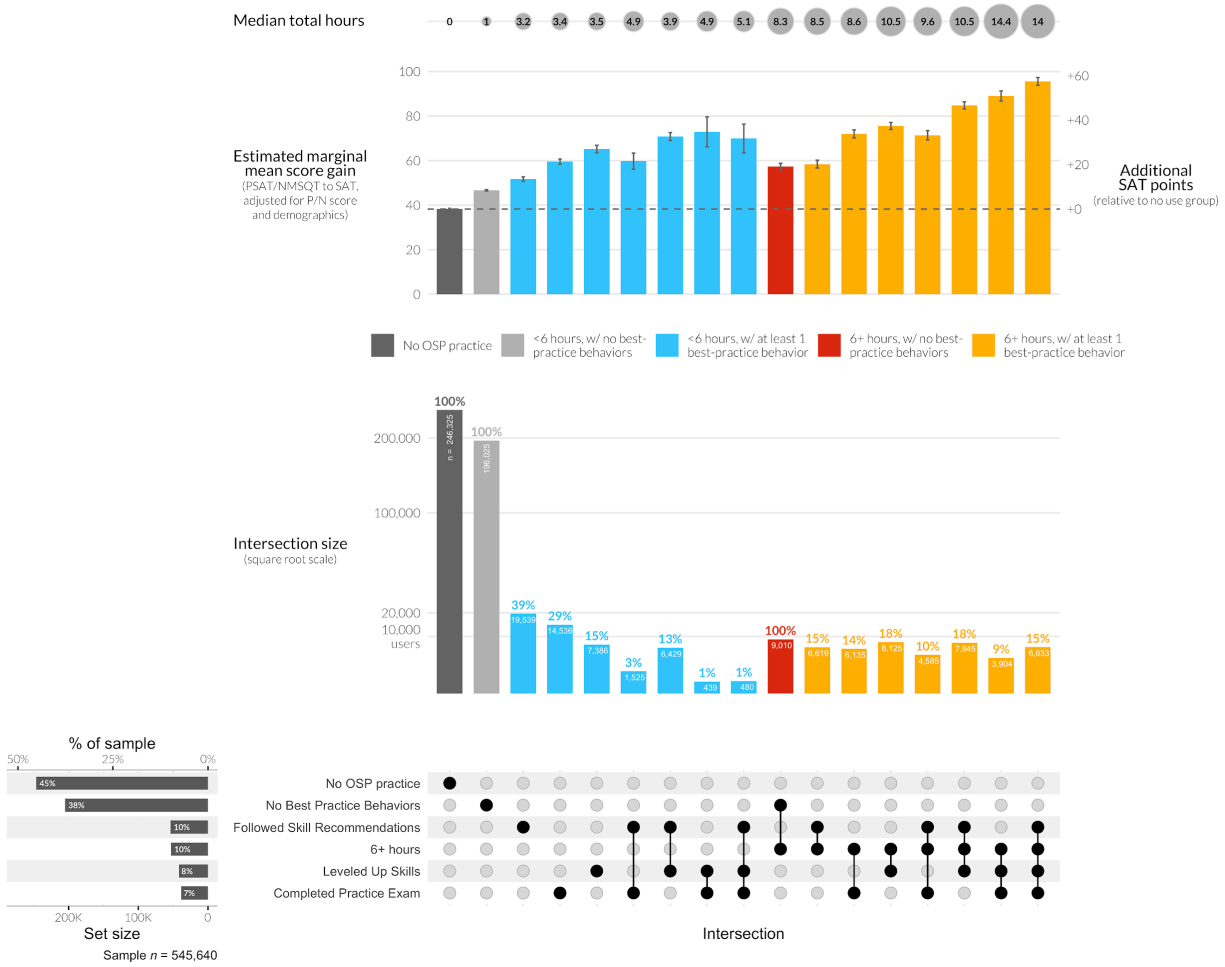


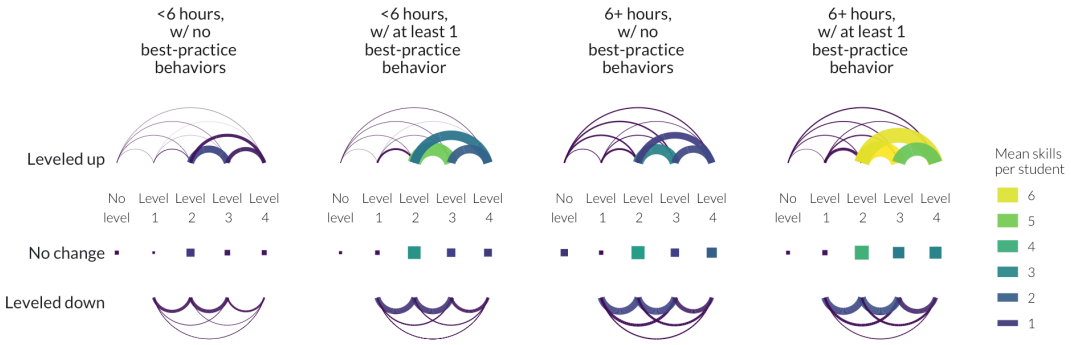
Figure 10. Intersection plot of both sample representation and estimated marginal mean score gain across usage groups showing different best practice combinations.

Leveling up skills deeper dive

In the first analysis in this subsection, we showed that students who engaged in at least one of the best practice behaviors showed better outcomes on the SAT than students who did not engage in one of these behaviors. Additionally, in the previous subsection, we discussed the interconnectedness of the best practices. One criticism of the best practice behavior, leveling up, is that it may stack the deck in favor of the best practice group. We used leveling up skills as an indication that students are engaging in the meaningful study and targeted practice necessary to improve performance. However, the measure itself is intrinsically a measure of performance—in order to level up, students need to answer problems correctly.

In this sense the variable may simply be picking up on a latent measure of general ability, and therefore it is not inherently surprising that students in the "best practice" group performed at higher levels.

The critique against leveling up skills is certainly valid. However, there are multiple reasons to suspect it is not unfairly biasing the best practice groups. Recall that in OSP, the difficulty of problems are scaled to the level of the learner based on PSAT item performance. Thus, leveling up should theoretically be just as viable for low-performing students as for high-performing ones. This can be observed in Figure 11, which shows in detail the descriptive data on leveling up skills across the four usage groups. The arches in the figure illustrate the movement of a skill from the initial level given to a student—the thicker the line the more skills were advanced from that level. The major takeaway in this figure is that leveling up happens in all of the groups. Even learners who spend a small amount of time on OSP and do not use a best practice are still able to level up in skills. Not surprisingly, the groups with one best practice behavior tend to level up more skills then the groups with no best practice behaviors. However, when we examine the table at the bottom of Figure 11, these students are also far more likely to attempt leveling up more skills. For example, looking only at students with six or more hours of OSP, students with one best practice behavior attempted a median 32 skills, whereas students with no best practice behaviors attempted a median 16 skills. Students in the best practice groups do appear to answer problems with greater accuracy, which would lead to more leveling up. However, it is difficult to determine whether this is a systematic bias or simply the product of learning via greater overall activity.



	<6 hours, w/ no best-practice behaviors	<6 hours, w/ at least 1 best-practice behavior	6+ hours, w/ no best-practice behaviors	6+ hours, w/ at least 1 best-practice behavior
Group n	196,025	50,334	9,010	43,946
N with skills leveled up	101,296	41,575	7,572	42,184
% with skills leveled up	52%	83%	84%	96%
Median problems attempted	27	159	291	460
Median problem accuracy	56.4%	63.0%	58.1%	67.0%
Median skills attempted	2	20	16	32
Median skills leveled up	1	9	6	17
Median skills leveled up to 4	0	2	2	8
Median attempts on leveled up skills	1.0	1.0	1.5	2.0

Figure 11. Skill level changes on OSP by each usage group. Lines denote the average (mean) number of skills per user.

2d. Are certain students more likely to engage in best practices?

Some students subgroups spend more time on OSP, but are slightly less likely to engage in best practices.

In this subsection, we examine whether there are group differences in students who are more or less likely to engage in best practice behaviors. A full breakdown of the models are in Appendix G. As with the above analyses (Figure 7), we focus on the most highly observed subgroups in each category. Figure 12 breaks down the likelihood of engaging in best practice behaviors and spending six plus hours by different student characteristics: gender, race, and parental education. It also shows the likelihood of best practice behavior against the number of hours on OSP and PSAT/NMSQT scores. We see that the greater time on OSP is related to more best practice behaviors and that students who score higher on the PSAT are more likely to engage in best practice behaviors. However, one important pattern demonstrated in this figure is that while some groups (e.g., Asian and Black or African American students) are more likely to spend at least six hours on OSP, that time is not necessarily spent engaging in best practices, which they are less likely to complete.

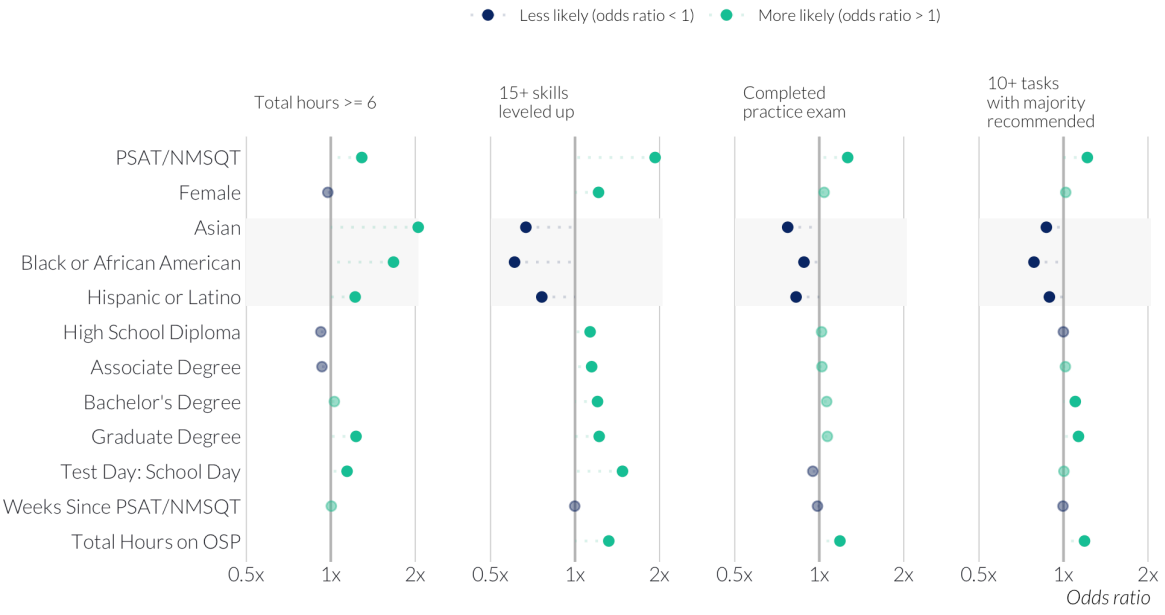


Figure 12. Odds of engaging in best practices by subgroups.

This pattern is an important potential difference in how these groups have engaged with OSP and suggests that further work is needed to understand why some groups are less likely to engage with these OSP features. However, it is also important to note that the empirical differences in these groups' best practice behaviors are small. In Table 2 from section 1a we show the percentages for target usage broken down by ethnicity. Although the visualization of the odds ratio above does show a statistically meaningful difference while holding time constant, the practical difference represents a very small percentage point gap between these groups.

Nevertheless, evaluating the impact of OSP for these student groups, particularly groups that are traditionally underserved in education, is an important focus of this report. While this signal is small, it does suggest that some students may not be as likely to utilize the best practice behaviors, which future work will continue to examine.

Discussion

In accordance with the call for more rigorous research on the effects of test preparation for college admission exams (Briggs, 2009), this report provides a comprehensive description of students who prepared for the revised SAT using a free online practice tool: OSP. Nearly 25% of the class of 2019 SAT test takers linked their Khan Academy and College Board accounts, and were demographically similar to the population of students who took a PSAT/NMSQT and the subsequent SAT. Students spent relatively little time on the platform, with 38% of students spending at least 1 hour or doing 50+ problems and only 10% practicing for six or more hours, consistent with other studies finding low rates of completion and high rates of dropoff for large, free online learning platforms (Gütl, Rizzardini, Chang & Morales, 2014; Kizilcec & Halawa, 2015). When students did practice, they did most of their practice in the two months leading up to taking the SAT.

Previous studies have associated the amount of time spent preparing with impacts on scores. One analysis found that each additional hour of tutoring was associated with an increase of 2.34 SAT points (Appelrouth et al., 2015) and another analysis found that 6 to 8 hours of OSP usage was associated with an additional 30-point score increase from PSAT/NMSQT to students' last SAT (College Board 2018b). The current analysis also shows a positive association between hours on OSP and SAT performance for composite SAT as well as math and ERW sections, although the majority of students spent 3 hours or less on OSP. Because time spent on OSP is only a high-level description of use, we also examined particular behaviors that were associated with better SAT outcomes.

This analysis demonstrates that practicing on OSP for at least six hours, along with one of the best practices, is associated with 39 additional points on the SAT composite score, an effect size of 0.20. This positive effect of test preparation is similar to the magnitude reported in studies of coaching classes and tutoring. Like previous studies, the 39-point difference is similar to other analyses of coaching associated with the SAT (Briggs, 2009; Montgomery & Lilly, 2012). This difference may have practical significance for colleges who use cut scores to make admissions decisions and other researchers have found that even small differences in SAT scores can have an impact on college options, particularly for lower-scoring students (Briggs, 2009; Goodman et al., 2017).

While there were no substantive differences by race or parental education in who used OSP, there were differences in the likelihood to use best practices. Asian students and Black students were more likely to practice for at least six hours, but somewhat less likely to do any of the best practices. This finding has direct product improvement implications. We are refreshing OSP to create an updated experience. As part of these updates we plan to guide the user to the best practice behaviors and to make it an integral part of the experience.

Limitations and Future Work

In this study, we presented a uniquely broad sample from the 2019 cohort of high school students taking the SAT, and we evaluated the impact of Official SAT Practice usage on students' SAT achievement in a real-world context. This research design allowed us to examine student behavior in its real, situated context and pointed toward broad patterns of benefit from both best practice behaviors and time spent on OSP.

However, there are some crucial limitations to consider for these findings. As highlighted throughout, when defining data from the OSP platform we were limited to primarily examining the time that users spent, along with whether users completed a few high-level definitions of best practice behaviors. We were unable to deeply examine differences in domain specificity for students' OSP use. It is possible that student usage patterns in domain concentration are important to understand the impact of OSP use on math and ERW outcomes. It is also possible that best practices differ between the math and ERW content on OSP, and that more holistic measures of student performance on OSP over time will prove important to evaluating OSP as an intervention. Best practice behaviors defined in this report are limited measures of student performance on Official SAT Practice. There are many other important measures to consider in student performance that could be explored on OSP, such as how students benefit from reviewing their work on OSP. Future work will build a deepened understanding of OSP usage and enable an examination of these and other usage questions.

This analysis focused solely on data from students who consented to data sharing between College Board and Khan Academy. While this data included important individual characteristics, such as parental education and prior achievement, we were unable to consider student motivation or contextual information. It is very likely that these differences are an important piece of understanding why some students engage with SAT preparation in different ways from other students. We were further unable to capture information about external study for the SAT, such as practice exams that students took that were not on the OSP platform, or if students used OSP with additional study resources like a private tutor or an SAT class. If some students show low activity on OSP, but also receive a high amount of SAT preparation elsewhere, this may mitigate the impact measured from the OSP platform.

Finally, as noted throughout, this study is an observational design. While this provides the ability to assess real-world behavior over a large cohort of students, our analysis cannot manipulate experimental groups, assign student behavior, or directly compare OSP usage against different SAT preparation methods. Our findings speak to the efficacy of OSP within a sample, as observed on those students' pre- and post-intervention test performance, but cannot control for students' self-selection and possible systematic confounds between our observed groups. Our findings also cannot directly compare how OSP performs against other SAT preparation materials or interventions, or against a control intervention.

Conclusion

This study provided new evidence on the use of a free, online tool to prepare for the revised SAT, contributing to the test preparation literature by providing necessary and called-for new evidence on the impact of digital preparation on the current SAT (Briggs, 2009). In the large sample of over half a million students, we show a positive association between time spent on OSP and SAT outcomes; results that remain positive across race and parental education. The way students use the platform matters. Our analyses show that students who practice six or more hours—along with best practices like leveling up skills, taking a practice test, and following the recommended practice—achieve higher SAT scores. Importantly, while the benefits above hold true across most student demographics, we see fewer lower-performing and underrepresented students, those we most seek to help, who are taking these actions, even when they spend the same amount of time on the platform. For this reason, the College Board and Khan Academy will work diligently with our partners across the country through programmatic supports and platform refinements in coming years to ensure that all students can follow these best practices. With this study, we provide the first specific guidance for how students should spend time on OSP and will continue to explore and refine our recommendations.

While the data associating best practices with score increases are promising, we need more research on implementation to ensure that when best practices are used more broadly, the associations remain as strong. Further research will help our understanding of student progress, any differences in adoption of best practice behaviors, and how supports such as school-day implementation and educator tools can help keep all students engaged and on track. Moreover, as more education moves online, we hope to learn more about how educational platforms are used together to support students during typical and uncertain times.

These findings provide a first step toward exploring the impact of the free, online SAT practice resource developed by Khan Academy and College Board. While more work is needed, there are many features within Official SAT Practice that have traditionally been difficult for learners to access outside of high-touch, expensive preparation products, such as diagnostic skill evaluations, which link directly to SAT content and subsequent recommendations. Future work on these comparisons could provide insight on both how to refine the design of these features and the impact of providing personalized intervention at scale to students who otherwise would not be able to access this kind of preparation.

References

Adesope, O., Trevisan, D., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. Review of Educational Research, 87(3): 659–701.

Appelrouth, J., Zabrucky, K., & Moore, D. (2015). Preparing students for college admissions tests. Assessment in Education: Principles, Policy & Practice, 24(1); 78–95.

Appelrouth, J., Moore, D., Zabrucky, K., & Cheung, J. (2018). Preparing for high-stakes admissions tests: A moderation mediation analysis. International Researcher in Higher Education, 3(3): 32–50.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. Chance, 14(1): 10–18.

Briggs, D. (2002). SAT Coaching, Bias and Causal Inference. Unpublished doctoral dissertation, University of California, Berkeley.

Briggs, D. C. (2009). Preparation for college admission exams (2009 NACAC Discussion Paper). Arlington, VA: National Association for College Admission Counseling.

Buchmann, C., Condron, D., & Roscigno, V. (2010). Shadow Education, American Style: Test Preparation, the SAT and College Enrollment. Social Forces, 89(2): 435–461.

Byun, S. & Park, H. (2012). The academic success of East Asian American youth: The role of shadow education. Sociology of Education, 85(1), 40–60.

Center for Research and Reform in Education. (2019). Evidence for ESSA. Retrieved from https://www.evidenceforessa.org/

College Board. (2018a). Unpublished data.

College Board. (2018b). Delivering Opportunities: SAT Suite of Assessments Results 2016-17. Retrieved from https://research.collegeboard.org/pdf/college-board-delivering-opportunities-sat-suite-results-2016-17.pdf.

College Board. (2019a). SAT Suite Results: Class of 2019. Retrieved from https://reports.collegeboard.org/sat-suite-program-results/class-2019-results.

College Board. (2019b). SAT Understanding Scores 2019. Retrieved from https://collegereadiness.collegeboard.org/pdf/understanding-sat-scores.pdf.

Goodman, J., Hurwitz, M., & Smith, J. (2017). Access to 4-Year Public Colleges and Degree Completion. Journal of Labor Economics, 35 (3): 829–867.

Guo, S. & Fraser, M. W. (2015). Propensity score analysis: Statistical methods and applications (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014). Attrition in MOOC: Lessons

learned from drop-out students. In: International workshop on learning technology for education in cloud (pp. 37–48).
Haimovitz, K., Shankar, P., Gallop, R., Yeager, D., Gross, J. J., & Duckworth, A. L. *Under review.* Strategic Self-Control Supports Studying for the SAT: Evidence From Three National Field Studies.
Institute of Education Sciences. (2019). *What works clearinghouse.* Retrieved from https://ies.ed.gov/ncee/wwc/
Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In Proceedings of the Second (2015) ACM Conference on Learning@ Scale (pp. 57–66).
Kraft, M. A. (2018). Interpreting effect sizes of educational interventions. *Brown University working paper.* Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_2018_interpreting_effect_sizes.pdf
Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). Handbook of test development. New York:Routledge.
Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010).Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. Washington DC: U.S. Department of Education. Retrieved from https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf
Montgomery, P., & Lilly, J. (2012). Systematic reviews of the effects of preparatory courses on university entrance examinations in high school-age students. International Journal of Social Welfare, 21(1), 3–12.
Moore, R., Sanchez, E., & San Pedro, S. (2019). College Entrance Exams: How does test preparation affect retest scores? Iowa City, IA: ACT.
Park, H. & Becks, A. (2015). Who benefits from SAT prep?: An examination of high school context and race/ethnicity. The Review of Higher Education, 39(1): 1–23.
Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin and Company.
Thoemmes, F., & Ong, A. D. (2016). A primer on inverse probability of treatment weighting and marginal structural models. Emerging Adulthood, 4(1), 40-59.
U.S. News and World Report. (2020). Find an SAT Tutor: Free SAT test prep options are available, but some parents may opt to hire a tutor to help their child study. Retrieved from https://www.usnews.com/education/find-sat-tutor, retrieved May 21, 2020.
What Works Clearinghouse. (2019). About us. Washington, DC: Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/WhatWeDo, retrieved May 31, 2019.

R References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
Clarke, E. and Sherrill-Mix, S. (2017). ggbeeswarm: Categorical Scatter (Violin Point) Plots. R package version 0.6.0. https://CRAN.R-project.org/package=ggbeeswarm
Greifer, N. (2020). WeightIt: Weighting for Covariate Balance in Observational Studies. R package version 0.9.0. https://CRAN.R-project.org/package=WeightIt
Karlsson, A. and Clements, M. (2018). biostat3: Utility Functions, Datasets and Extended Examples for Survival Analysis. R package version 0.1.3. https://CRAN.R-project.org/package=biostat3
Kassambara, A. (2018). ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'. R package version 0.1.2. https://CRAN.R-project.org/package=ggcorrplot
Fox, J. (2003). Effect Displays in R for Generalised Linear Models. Journal of Statistical Software, 8(15), 1-27. URL http://www.jstatsoft.org/v08/i15/.
Henry, L. and Wickham, H. (2019). purrr: Functional Programming Tools. R package version 0.3.2. https://CRAN.R-project.org/package=purrr

Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.3. https://CRAN.R-project.org/package=emmeans
Lüdecke, D. (2018). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.6.2, https://CRAN.R-project.org/package=sjPlot. doi: 10.5281/zenodo.1308157.
Lumley, T (2020) "survey: analysis of complex survey samples". R package version 4.0.
R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Robinson, D. and Hayes, A. (2019). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.2. https://CRAN.R-project.org/package=broom
Rudis, B. (2019). hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'. R package version 0.6.0. https://CRAN.R-project.org/package=hrbrthemes
Slowikowski, K. (2018). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=ggrepel
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
Wickham, H. and Bryan, J. (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl
Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse
Wickham, H. (2019). bigrquery: An Interface to Google's 'BigQuery' 'API'. R package version 1.1.0. https://CRAN.R-project.org/package=bigrquery
Wilke, C. O. (2018). ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.1. https://CRAN.R-project.org/package=ggridges
Wilke, C. O. (2019). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 0.9.4. https://CRAN.R-project.org/package=cowplot
Xie, Y. (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.22.
Zhu, H. (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra

Appendix A. Table of Variables

Table A1.
Summary of Analysis Variables.

	Variable	Definition
Primary outcome	SAT composite score	Sum of the Math and Evidence-based Reading and Writing sections of students' first SAT. Range from 400–1600. Note: Excludes performance on the optional Essay section.
Usage variables for Official SAT Practice (*curated to only include usage between each student's PSAT/NMSQT and first SAT dates*)	Time on Official SAT Practice	Time is analyzed either as a continuous measure (hours) or dichotomized to indicate whether a student studied for at least six hours, depending on the research question.
	Completed a full-length practice exam	Boolean—True if the student completed all four sections of a full-length practice exam on Official SAT Practice. Note: does not count practice exams that were downloaded and completed on paper.
	Leveled up 15+ skills through practice	Boolean —True if the student's last observed level on 15+ skills was higher than their starting level on those skills during the interval between their PSAT/NMSQT and first SAT. Note: If a student levels up on a skill but then levels back down, the net is *no change* and is not counted toward the threshold of 15+ skills leveled up.
	Followed skill practice recommendations on a majority of 10+ tasks	Boolean—True if the student completed 10+ skill practice tasks (e.g., practice exercises, timed mini-sections) and >50% of those tasks were from their personalized practice recommendation queue.
Statistical controls	PSAT/NMSQT composite score	Continuous value in the range 320–1520, (grand mean centered).
	Gender	Factor with 2 levels: ● Female ● Male **(reference category)**
	Race/Ethnicity	Dummy codes for 8 categories: ● American Indian or Alaska Native ● Asian ● Black or African American ● Hispanic or Latino ● Native Hawaiian or Other Pacific Islander ● Two or more races ● White **(reference category)** ● *No response*
	Parental education level	Categorical variable with five levels: ● No high school diploma **(reference category)** ● High school diploma or equivalent OR Business or trade school ● Associate or two-year degree ● Bachelor's or four-year degree ● Graduate or professional degree ● *No response*
	SAT Test Day	Categorical variable denoting whether the student's SAT occurred on a weekend or weekday.
	Weeks since PSAT/NMSQT	Integer indicating the number of calendar weeks between when the student took the PSAT/NMSQT and when they took the SAT for the first time (mean centered).

Appendix B. Modeling the Relationship Between Time Using OSP and SAT Achievement

In this section of the Appendix, we provide a technical breakdown of the analysis presented in Section 2a of the report. We recommend that the reader reviews that section of the report for context and rationale before reading this section.

We estimated the effect of the amount of time using OSP on SAT performance using sequenced multiple linear regression. In each step of the sequence, we added a variable or set of variables in order of causal priority, starting with PSAT/NMSQT and concluding with OSP usage. There were four steps in total, specified by the following four models:

(1) $SAT_i = \beta_0 + \beta_1 PSAT_i + \varepsilon_i$

(2) $SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \varepsilon_i$

(3) $SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_{i+\varepsilon_i}$

(4) $SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 OSP\ hours_i + \beta_8 OSP\ hours_i{}^2 + \varepsilon_i$

In these equations, SAT was either the composite, math, or ERW SAT scores. The $PSAT/NMSQT$ variable was student either the composite, math, or ERW PSAT/NMSQT score. The PSAT/NMSQT scored used depended on the SAT score being predicted. For example, if the composite SAT was being predicted, then the composite PSAT/NMSQT was used. PSAT/NMSQT scores were grand mean centered. $Gender$ and $Ethnicity$ are self-explanatory demographic variables and were dummy coded for the model. $Parental\ Education$ referred to the highest level of education achieved by the child's parents. $Test\ Day$ refers to whether students took the exam on a "weekend" or "weekday." $Weeks\ Since\ PSAT$ was the number in weeks that elapsed between taking the PSAT/NMSQT and the SAT, and was grand mean centered. A full description of variables are shown in Appendix A.

For the present analysis, $OSP\ hours$ and $OSP\ hours^2$ were the critical dependent variables, and refer to the number of hours that a student spent using OSP. This variable was specified as a second-order polynomial (quadratic term) to allow for the possibility of diminishing returns over time. That is, while we hypothesize a positive association between studying on Official SAT Practice and student's SAT achievement, we do not expect the form of this association to be a monotonic increase in perpetuity. Rather, we anticipate that the benefits of studying on Official SAT Practice might begin to diminish after a certain point, if for no other reason than that the practice content is expansive but ultimately finite, meaning that students will eventually exhaust the unique practice resources. Note that $OSP\ hours$ was extremely skewed, with some extreme values on the high end of the distribution. We decided to right censor this variable, replacing any values larger than 30 with 30. We did not center $OSP\ hours$ in order to preserve its meaningful zero point. The paraments β_0 and ε_i refer to the intercept and error, respectively.

Results from the models predicting composite, math, and ERW are shown at the end of this Appendix on Tables B1, B2, and B3, respectively. 95% confidence intervals around the estimates are also shown. We urge readers to use caution when comparing the coefficient estimates across tables. In particular, the composite SAT is on a different scale than math and ERW, given that the composite scores are the sum of the math and ERW. For instance, a variable that produces a larger coefficient estimate for composite than math or ERW is not necessarily meaningful. Coefficient estimates between math and ERW are directly

comparable. Also note that due to the large sample size, practically all coefficients reach the threshold of statistical significance. For this reason, we encourage readers to focus more on interpreting the magnitude of the effects, as well as the precision around the estimates provided by the 95% confidence intervals.

There are several effects present on Tables B1, B2, and B3 that are worth discussing. First is the strong correlation between PSAT/NMSQT performance and SAT performance. The PSAT/NMSQT coefficients can be interpreted as the expected increase in SAT for every additional PSAT/NMSQT point. For model 1 in each table, we see a very strong relationship between PSAT/NMSQT and SAT, nearing almost a 1:1 relationship. The extremely low uncertainty around the effect of PSAT/NMSQT is also remarkable, as the 95% confidence intervals practically converge to the point estimate of the effect. A final point worth mentioning regarding PSAT/NMSQT is the high R^2values observed on Model 1 of each table. PSAT/NMSQT alone explains 85%, 77%, and 79% of the variance for composite, math, and ERW scores, respectively. These are extremely high values for social science research. The sum total of these findings speak to the reliability of the PSAT/NMSQT at predicting SAT performance, and the central challenge in "moving the needle" above and beyond a student's prior achievement.

We will only briefly discuss the demographic variables and test characteristics, which were added as sets in models 2 and 3, respectively. In general, we observed typical score patterns reported by the College Board. The observed effects of test day are most likely an artifact of self-selection bias—notably, students who take the exam during the weekend are most likely to be motivated, high-performing students. There were interesting effects of weeks since the PSAT/NMSQT—notably the trend was negative for math, but positive for the composite and ERW. It is worth noting that the inclusion of additional variables in each step provided only extremely small improvements in R^2above and beyond that of model 1. Although these improvements to R^2were small, AIC was minimized at each step, suggesting that the added model complexity was warranted (AIC cite).

Lastly, we will discuss the results of model 4, which added the effect of $OSP\ hours$to model 3. Additional discussion can be found in the main body of the text. Because we included $OSP\ hours$ as a second order polynomial, interpretation of the coefficients is not as straightforward as the other variables. Nevertheless, the first order term indicates a general positive relationship between $OSP\ hours$and SAT performance for all three SAT outcomes. The fact that the second order $OSP\ hours^2$term was reliably negative and different from 0 indicates that there is a downward facing curvilinear relationship between $OSP\ hours$and SAT performance. We refer the reader to Section 2a of the main text for a more nuanced discussion of the effects of time, as well as a visual representation. We note that, like the demographic and test characteristic variables, the addition of $OSP\ time$ provided only small improvements to R^2above and beyond model 3, suggesting these variables provide only marginal benefits for prediction. But note that the reduction in AIC estimates do support the inclusion of the variable in the model. Regardless, as the nature of this inquiry is causal in nature, the estimated regression coefficients are of primary importance.

Note that in this section, we express the results as both in terms of the estimated SAT points and as an effect size. The effect sizes were calculated using the predictions of Model 4 discussed above, specifically as:

$$\frac{\hat{Y}_H - \hat{Y}_{H=0}}{SD(Y)}$$

Where $\hat{Y}$ is the predicted SAT score, $H$is the number of hours spent on OSP, and $SD(Y)$is the standard deviation of all observed SAT scores.

Table B1.
Linear Regression Estimates of Composite Achievement on First SAT.

	Model 1		*Model 2*		*Model 3*		*Model 4*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*
Intercept	1105.73	1105.53 – 1105.92	1102.13	1101.30 – 1102.95	1107.12	1106.26 – 1107.98	1101.57	1100.71 – 1102.42
PSAT/NMSQT (Composite)	0.95	0.95 – 0.95	0.92	0.92 – 0.92	0.92	0.92 – 0.92	0.91	0.91 – 0.91
Gender: Female			-6.84	-7.23 – -6.44	-7.18	-7.57 – -6.78	-7.39	-7.78 – -7.00
Ethnicity: American Indian			-17.71	-20.91 – -14.51	-16.97	-20.16 – -13.77	-17.41	-20.56 – -14.25
Ethnicity: Asian			9.42	8.75 – 10.09	8.71	8.04 – 9.38	5.16	4.49 – 5.83
Ethnicity: Black			-14.93	-15.61 – -14.25	-15.08	-15.76 – -14.41	-17.14	-17.81 – -16.48
Ethnicity: Hispanic/Latinx			-10.46	-10.98 – -9.93	-10.93	-11.46 – -10.41	-11.56	-12.08 – -11.04
Ethnicity: Native Hawaiian/Pac. Islander			-11.89	-16.47 – -7.31	-12.98	-17.56 – -8.41	-13.38	-17.90 – -8.86
Ethnicity: Unknown			-0.92	-2.51 – 0.67	-1.13	-2.72 – 0.45	-3.04	-4.61 – -1.47
Ethnicity: Two or more			-2.06	-3.02 – -1.10	-2.83	-3.80 – -1.87	-3.51	-4.46 – -2.55
Parental Education: High School Diploma			2.49	1.63 – 3.35	2.27	1.41 – 3.12	2.59	1.75 – 3.43
Parental Education: Associate Degree			5.7	4.88 – 6.52	5.2	4.38 – 6.02	5.52	4.71 – 6.33
Parental Education: Bachelor's Degree			14.41	13.60 – 15.22	13.7	12.89 – 14.51	13.69	12.88 – 14.49
Parental Education: Graduate Degree			22.25	21.39 – 23.10	21.3	20.44 – 22.16	20.7	19.85 – 21.54
Parental Education: No response			-1.88	-3.40 – -0.36	-1.79	-3.30 – -0.27	-1.53	-3.03 – -0.03
Test Day: School Day					-7.44	-7.86 – -7.02	-8.74	-9.15 – -8.33
Weeks since PSAT					0.19	0.17 – 0.20	0.11	0.09 – 0.12
OSP hours							3.94	3.83 – 4.04
OSP hours^2							-0.06	-0.07 – -0.06
Observations	545640		545640		545640		545640	
R^2 / R^2 adjusted	0.853 / 0.853		0.857 / 0.857		0.857 / 0.857		0.861 / 0.861	
AIC	6242038.447		6229800.463		6227901.087		6214917.716	

Table B2.
Linear Regression Estimates of Math Achievement on First SAT

	Model 1		*Model 2*		*Model 3*		*Model 4*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*
Intercept	549.04	548.90 – 549.17	545.25	544.69 – 545.82	549.2	548.61 – 549.79	545.99	545.41 – 546.58
PSAT/NMSQT (Math)	0.95	0.94 – 0.95	0.89	0.89 – 0.89	0.89	0.88 – 0.89	0.88	0.88 – 0.88
Gender: Female			-5.25	-5.52 – -4.98	-5.69	-5.96 – -5.42	-5.9	-6.16 – -5.63
Ethnicity: American Indian			-15.08	-17.28 – -12.89	-14.31	-16.50 – -12.12	-14.46	-16.63 – -12.29
Ethnicity: Asian			10.38	9.92 – 10.84	9.5	9.04 – 9.96	7.42	6.96 – 7.88
Ethnicity: Black			-14.69	-15.15 – -14.23	-14.6	-15.06 – -14.13	-15.75	-16.21 – -15.29
Ethnicity: Hispanic/Latinx			-7.98	-8.34 – -7.62	-8.25	-8.61 – -7.89	-8.58	-8.93 – -8.22
Ethnicity: Native Hawaiian/Pac. Islander			-6.93	-10.07 – -3.79	-7.61	-10.75 – -4.48	-7.78	-10.89 – -4.67
Ethnicity: Unknown			-0.32	-1.41 – 0.77	-0.43	-1.52 – 0.65	-1.57	-2.64 – -0.49
Ethnicity: Two or more			-1.25	-1.91 – -0.59	-1.69	-2.35 – -1.03	-2.1	-2.76 – -1.45
Parental Education: High School Diploma			2.76	2.17 – 3.35	2.61	2.03 – 3.20	2.75	2.17 – 3.33
Parental Education: Associate Degree			5.68	5.12 – 6.24	5.27	4.71 – 5.83	5.38	4.83 – 5.93
Parental Education: Bachelor's Degree			12.58	12.02 – 13.13	11.62	11.06 – 12.17	11.47	10.92 – 12.02
Parental Education: Graduate Degree			17.84	17.26 – 18.43	16.46	15.88 – 17.04	15.92	15.34 – 16.49
Parental Education: No response			-2.33	-3.37 – -1.28	-2.36	-3.40 – -1.32	-2.21	-3.24 – -1.18
Test Day: School Day					-7.26	-7.55 – -6.98	-7.99	-8.28 – -7.71
Weeks since PSAT					-0.02	-0.04 – -0.01	-0.07	-0.08 – -0.06
OSP hours							2.31	2.24 – 2.39
OSP hours^2							-0.03	-0.04 – -0.03

	Model 1	*Model 2*	*Model 3*	*Model 4*
Observations	545640	545640	545640	545640
R_2 / R_2 adjusted	0.772 / 0.772	0.780 / 0.780	0.781 / 0.781	0.785 / 0.785
AIC	5837375.642	5818889.62	5816388.643	5806169.893

Table B3
Linear Regression Estimates of ERW Achievement on First SAT

	Model 1		*Model 2*		*Model 3*		*Model 4*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*
Intercept	556.69	556.58 – 556.81	552.76	552.26 – 553.25	555.16	554.65 – 555.67	552.42	551.91 – 552.94
PSAT/NMSQT (ERW)	0.88	0.87 – 0.88	0.84	0.84 – 0.84	0.84	0.84 – 0.84	0.83	0.83 – 0.83
Gender: Female			-1.44	-1.67 – -1.20	-1.59	-1.82 – -1.35	-1.62	-1.85 – -1.39
Ethnicity: American Indian			-10.29	-12.20 – -8.38	-9.92	-11.82 – -8.01	-10.1	-11.99 – -8.20
Ethnicity: Asian			1.08	0.68 – 1.48	0.63	0.23 – 1.03	-1.11	-1.51 – -0.71
Ethnicity: Black			-7.64	-8.04 – -7.23	-7.69	-8.09 – -7.28	-8.6	-9.00 – -8.20
Ethnicity: Hispanic/Latinx			-6.09	-6.41 – -5.78	-6.32	-6.63 – -6.00	-6.58	-6.89 – -6.27
Ethnicity: Native Hawaiian/Pac. Islander			-9.7	-12.44 – -6.97	-10.22	-12.95 – -7.48	-10.37	-13.09 – -7.66
Ethnicity: Unknown			-2.11	-3.06 – -1.16	-2.21	-3.16 – -1.26	-3.09	-4.03 – -2.15
Ethnicity: Two or more			-1.15	-1.72 – -0.57	-1.5	-2.07 – -0.92	-1.81	-2.38 – -1.24
Parental Education: High School Diploma			1.74	1.23 – 2.25	1.65	1.14 – 2.16	1.81	1.31 – 2.32
Parental Education: Associate Degree			3.52	3.03 – 4.01	3.3	2.81 – 3.79	3.46	2.97 – 3.94
Parental Education: Bachelor's Degree			9.06	8.57 – 9.54	8.68	8.19 – 9.16	8.67	8.19 – 9.15
Parental Education: Graduate Degree			14.74	14.23 – 15.26	14.22	13.70 – 14.73	13.92	13.41 – 14.43
Parental Education: No response			-0.47	-1.38 – 0.44	-0.43	-1.34 – 0.48	-0.28	-1.18 – 0.62
Test Day: School Day					-3.85	-4.09 – -3.60	-4.45	-4.70 – -4.20
Weeks since PSAT					0.07	0.06 – 0.08	0.03	0.02 – 0.04
OSP hours							1.98	1.91 – 2.04
OSP hours^2							-0.04	-0.04 – -0.04
Observations	545640		545640		545640		545640	
R_2 / R_2 adjusted	0.794 / 0.794		0.798 / 0.798		0.799 / 0.799		0.802 / 0.802	
AIC	5678548.585		5667551.176		5666356.059		5658481.889	

Appendix C. Modeling the Relationship Between Time Using OSP and SAT Achievement As a Function of Student Characteristics

In this Appendix, we provide a technical breakdown of the analysis presented in Question 2b of the report. We recommend that the reader reviews that section of the report for context and rationale before reading this Appendix.

The goal of this analysis was to test whether the relationship between the amount of time spent using OSP and SAT performance interacted with student characteristics, specifically, gender, ethnicity, parental education, and PSAT/NMSQT. For each interaction, we first fit a base model using only the main terms. Then we fit a second model that included interaction terms. This procedure was repeated for each gender, ethnicity, parental education, and PSAT/NMSQT. As an example, the model specifications for testing the gender X OSP hours interaction are shown below:

$$(1)\ SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 OSP\ hours_i + \beta_8 OSP\ hours_i{}^2 + \varepsilon_i$$

$$(2)\ SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 OSP\ hours_i + \beta_8 OSP\ hours_i{}^2 + \beta_9 (OSP\ hours_i\ x\ Gender_i) + \beta_{10} (OSP\ hours_i{}^2\ x\ Gender_i) + \varepsilon_i$$

There are no terms introduced in these analyses that have not already been described previously, so we refer the reader to Appendixes A and B for further clarification. The interaction coefficients (β_9,β_{10}) represent the difference in the influence of the OSP hours for each level of the demographic variable against the reference level (i.e, the overall slope). For the quadratic term, it represents the difference between the groups in terms of how quickly the benefits of using OSP taper off (i.e, the curve of the relationship). If the 95% CI intervals of the estimates do not contain 0, then it suggests that the group level differs from the reference group along that dimension. For parsimony, we are only reporting coefficient estimates for the main effect terms and interaction terms.

The results for the interactions of gender, ethnicity, parental education, and PSAT/NMSQT and OSP hours are shown on Tables C1, C2, C3, and C4, respectively. In general, all interaction models provided an improved fit relative to the main effect models, based on the reduction in AIC. However, we can tell from the R^2Values that these were incredibly modest effects. In none of the interaction models were the changes in R^2even observable when rounding to three digits. When we examine the coefficients of the interaction terms, there are several interactions, but all are very small. For example, Table C1 shows that there was reliable interaction between gender and the quadratic of OSP Hours, indicating that the benefits of OSP Hours may taper off more rapidly for females than males. However, even if this effect is real, it is so small as to not be practically meaningful. As we noted in the main text, at six hours of usage, males and females were obtaining roughly the same benefits of OSP usage. The fact that the models were able to detect statistically significant effects of these interactions was not surprising given the massive size of this data set. For this reason, we err toward focusing on the practical significance of the results over the statistical significance. To this end, we do not see meaningful differences in benefits derived from the use of OSP across any of the categories examined.

Table C1.
Regression results for Gender X OSP Hours Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Gender: Female	-7.39 ***	-7.78 – -7.00	-7.04 ***	-7.51 – -6.57
OSP hours	3.94 ***	3.83 – 4.04	4.00 ***	3.83 – 4.16
OSP hours^2	-0.06 ***	-0.07 – -0.06	-0.06 ***	-0.07 – -0.05
Female x OSP hours			-0.09	-0.30 – 0.13
Female x OSP hours^2			-0.01	-0.02 – 0.00
Observations	545640		545640	
R^2 / R^2 adjusted	0.861 / 0.861		0.861 / 0.861	
AIC	6214917.716		6214893.28	
			** p<0.05 ** p<0.01 *** p<0.001*	

Table C2.
Regression results for Ethnicity X OSP Hours Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Ethnicity: American Indian	-17.41 ***	-20.56 – -14.25	-13.90 ***	-17.73 – -10.08
Ethnicity: Asian	5.16 ***	4.49 – 5.83	7.41 ***	6.59 – 8.23
Ethnicity: Black	-17.14 ***	-17.81 – -16.48	-15.12 ***	-15.93 – -14.31
Ethnicity: Hispanic/Latinx	-11.56 ***	-12.08 – -11.04	-10.60 ***	-11.21 – -10.00
Ethnicity: Native Hawaiian/Pac. Islander	-13.38 ***	-17.90 – -8.86	-10.96 ***	-16.40 – -5.52
Ethnicity: Unknown	-3.04 ***	-4.61 – -1.47	-1.73	-3.64 – 0.18
Ethnicity: Two or more	-3.51 ***	-4.46 – -2.55	-3.14 ***	-4.29 – -1.98
OSP hours	3.94 ***	3.83 – 4.04	4.51 ***	4.34 – 4.68
OSP hours^2	-0.06 ***	-0.07 – -0.06	-0.08 ***	-0.09 – -0.07
American Indian x OSP hours			-2.65 **	-4.63 – -0.68
Asian x OSP hours			-1.34 ***	-1.66 – -1.02
Black x OSP hours			-1.45 ***	-1.80 – -1.10
Hispanic x OSP hours			-0.73 ***	-1.01 – -0.46
Native Hawaiian/Pac. Islander x OSP hours			-1.76	-4.54 – 1.02
Unknown ethnicity x OSP hours			-0.79	-1.59 – 0.01
Two or more races x OSP hours			-0.27	-0.80 – 0.26
American Indian x OSP hours^2			0.07	-0.03 – 0.17
Asian x OSP hours^2			0.04 ***	0.02 – 0.05
Black x OSP hours^2			0.05 ***	0.03 – 0.06
Hispanic x OSP hours^2			0.02 ***	0.01 – 0.04
Native Hawaiian/Pac. Islander x OSP hours^2			0.04	-0.09 – 0.17
Unknown ethnicity x OSP hours^2			0.02	-0.02 – 0.05
Two or more races x OSP hours^2			0.01	-0.02 – 0.03
Observations	545640		545640	
R_2 / R_2 adjusted	0.861 / 0.861		0.861 / 0.861	
AIC	6214917.716		6214786.756	

** p<0.05 ** p<0.01 *** p<0.001*

Table C3.
Regression results for Parental Education X OSP Hours Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Parental Education: High School Diploma	2.59 ***	1.75 – 3.43	2.42 ***	1.40 – 3.43
Parental Education: Associate Degree	5.52 ***	4.71 – 6.33	5.12 ***	4.15 – 6.09
Parental Education: Bachelor's Degree	13.69 ***	12.88 – 14.49	13.09 ***	12.14 – 14.04
Parental Education: Graduate Degree	20.70 ***	19.85 – 21.54	20.61 ***	19.62 – 21.61
Parental Education: No response	-1.53 *	-3.03 – -0.03	-0.18	-1.97 – 1.61
OSP hours	3.94 ***	3.83 – 4.04	3.66 ***	3.28 – 4.03
OSP hours^2	-0.06 ***	-0.07 – -0.06	-0.05 ***	-0.07 – -0.03
High School Diploma x OSP hours			0.12	-0.36 – 0.61
Associate Degree x OSP hours			0.44	-0.01 – 0.90
Bachelor's Degree x OSP hours			0.52 *	0.10 – 0.94
Graduate Degree x OSP hours			0.15	-0.28 – 0.58
No response x OSP hours			-1.18 **	-2.04 – -0.32
High School Diploma x OSP hours^2			0	-0.03 – 0.02
Associate Degree x OSP hours^2			-0.02 *	-0.05 – -0.00
Bachelor's Degree x OSP hours^2			-0.02 *	-0.04 – -0.00
Graduate Degree x OSP hours^2			-0.01	-0.03 – 0.01
No response x OSP hours^2			0.04 *	0.00 – 0.09
Observations	545640		545640	
R^2 / R^2 adjusted	0.861 / 0.861		0.861 / 0.861	
AIC	6214917.716		6214906.814	

** p<0.05 ** p<0.01 *** p<0.001*

Table C4.
Regression results for PSAT/NMSQT X OSP Hours Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
PSAT/NMSQT (Composite)	0.91 ***	0.91 – 0.91	0.90 ***	0.90 – 0.91
OSP hours	3.94 ***	3.83 – 4.04	3.85 ***	3.74 – 3.96
OSP hours^2	-0.06 ***	-0.07 – -0.06	-0.06 ***	-0.07 – -0.06
PSAT/NMSQT x OSP hours			0.00 ***	0.00 – 0.00
PSAT/NMSQT x OSP hours^2			-0.00 ***	-0.00 – -0.00
Observations	545640		545640	
R_2 / R_2 adjusted	0.861 / 0.861		0.861 / 0.861	
AIC	6214917.716		6214716.474	

** p<0.05 ** p<0.01 *** p<0.001*

Appendix D. Modeling the Relationship Between Best Practice Behaviors on OSP and SAT Performance

In this section of the Appendix, we provide a technical breakdown of the analysis presented in Section 2c of the report. We recommend that the reader reviews that section of the report for context and rationale before reading this section.

We estimated the overall effect of engaging in best practice behaviors on OSP using sequenced multiple linear regression. In each step of the sequence, we added a variable or set of variables in order of causal priority, starting with PSAT, and concluding with the OSP usage. There were four steps in total, specified by the following four models:

(1) $SAT_i = \beta_0 + \beta_1 PSAT_i + \varepsilon_i$

(2) $SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \varepsilon_i$

(3) $SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \varepsilon_i$

(4) $SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 OSP\ Usage\ Group_i + \varepsilon_i$

In these equations, SAT was the composite SAT scores. The $PSAT$ variable was the grand mean centered composite PSAT/NMSQT score. $Gender$ and $Ethnicity$ are self-explanatory demographic variables and were dummy coded for the model. "Male" was the reference group for $Gender$ and "White" was the reference group for $Ethnicity$. $Parental\ Education$ referred to the highest level of education achieved by the child's parents, and was dummy coded with "Grade School" as the reference level. $Test\ Day$ refers to whether students took the exam on a "weekend" or "weekday," and was dummy coded with "weekday" as the reference level. $Weeks\ Since\ PSAT$ is the number in weeks that elapsed between taking the PSAT/NMSQT and the SAT. Finally, the $OSP\ Usage\ Group$ was a dummy coded categorical variable with 5 groups; (1) No OSP Usage, (2) less than six hours OSP, and no best practice behaviors, (3) less than six hours OSP, at least one best practice behavior, (4) six or more hours OSP, at least one best practice behavior, and (5) six or more hours OSP, no best practice behaviors. "No OSP Usage" was used as the reference level. The paraments β_0 and ε_i refer to the intercept and error, respectively.

Model results are shown at the end of this Appendix on Table D1, and 95% confidence intervals around the estimates are also shown. Note that models 1–3 results are identical to those from the analysis in Appendix B. Discussion of the OSP usage variables can be found in Section 2c. Inclusion of the OSP usage variable in Model 4 provided a reduction in AIC over Model 3, suggesting that the added complexity was warranted. Note that the change in R^2 was very small, meaning the added OSP usage variables did not contribute much in explaining overall variance in SAT performance above and beyond that of the PSAT/NMSQT scores. This is not particularly surprising, as the PSAT/NMSQT already accounts for so much of the variance.

Relative contribution of best practice behaviors

In order to estimate the relative effectiveness of each of the best practice behaviors, we fit the data to the following model:

$$(5)\ SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 Six\ hours\ OSP_i + \beta_8 Leveled\ Up_i + \beta_9 Practice\ Exam_i + \beta_{10} Followed\ Recommended\ Practice_i + \varepsilon_i$$

$Six\ hours\ OSP$was a dummy variable indicating whether a student had exceeded six hours using OSP. $Leveled\ Up$, $Practice\ Exam$, and $Followed\ Recommended\ Practice$were each dummy variables indicating whether a student met the threshold for each of the best practice behaviors (leveling up skills, completing a full-length practice exam, and following recommended practice). The full model results are shown in Table D2.

Table D1.
Linear Regression Estimates of Composite PSAT/NMSQT Achievement on First SAT

	Model 1		*Model 2*		*Model 3*		*Model 4*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*
Intercept	1105.73	1105.53 – 1105.92	1102.13	1101.30 – 1102.95	1107.12	1106.26 – 1107.98	1099.76	1098.89 – 1100.63
PSAT/NMSQT (Composite)	0.95	0.95 – 0.95	0.92	0.92 – 0.92	0.92	0.92 – 0.92	0.91	0.91 – 0.91
Gender: Female			-6.84	-7.23 – -6.44	-7.18	-7.57 – -6.78	-7.62	-8.01 – -7.23
Ethnicity: American Indian			-17.71	-20.91 – -14.51	-16.97	-20.16 – -13.77	-17.28	-20.44 – -14.13
Ethnicity: Asian			9.42	8.75 – 10.09	8.71	8.04 – 9.38	6.29	5.62 – 6.96
Ethnicity: Black			-14.93	-15.61 – -14.25	-15.08	-15.76 – -14.41	-16.32	-16.99 – -15.65
Ethnicity: Hispanic/Latinx			-10.46	-10.98 – -9.93	-10.93	-11.46 – -10.41	-11	-11.52 – -10.49
Ethnicity: Native Hawaiian/Pac. Islander			-11.89	-16.47 – -7.31	-12.98	-17.56 – -8.41	-13.12	-17.65 – -8.60
Ethnicity: Unknown			-0.92	-2.51 – 0.67	-1.13	-2.72 – 0.45	-2.49	-4.06 – -0.92
Ethnicity: Two or more			-2.06	-3.02 – -1.10	-2.83	-3.80 – -1.87	-3.22	-4.17 – -2.27
Parental Education: High School Diploma			2.49	1.63 – 3.35	2.27	1.41 – 3.12	2.47	1.63 – 3.32
Parental Education: Associate Degree			5.7	4.88 – 6.52	5.2	4.38 – 6.02	5.39	4.59 – 6.20
Parental Education: Bachelor's Degree			14.41	13.60 – 15.22	13.7	12.89 – 14.51	13.68	12.87 – 14.48
Parental Education: Graduate Degree			22.25	21.39 – 23.10	21.3	20.44 – 22.16	20.84	19.99 – 21.69

	Model 1		*Model 2*		*Model 3*		*Model 4*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*	*Estimates*	*CI*
Parental Education: No response			-1.88	-3.40 – -0.36	-1.79	-3.30 – -0.27	-1.62	-3.12 – -0.12
Test Day: School Day					-7.44	-7.86 – -7.02	-8.95	-9.36 – -8.53
Weeks since PSAT					0.19	0.17 – 0.20	0.1	0.08 – 0.12
OSP Usage: <6 hrs, w/ no best-practice behaviors							8.13	7.70 – 8.56
OSP Usage: <6 hrs,w/ at least 1 best-practice behavior							20.39	19.69 – 21.08
OSP Usage: 6+ hrs, w/ no best-practice behaviors							18.83	17.31 – 20.35
OSP Usage: 6+ hrs, w/ at least 1 best-practice behavior							39.22	38.48 – 39.96
Observations	545640		545640		545640		545640	
R_2 / R_2 adjusted	0.853 / 0.853		0.857 / 0.857		0.857 / 0.857		0.860 / 0.860	
AIC	6242038.447		6229800.463		6227901.087		6215424.782	

Table D2.
Regression Table of Relative Contribution Analysis

	Model 1	
Predictors	*Estimates*	*CI*
Intercept	1105.78	1104.43 – 1107.13
Six hours of OSP	10.79	9.95 – 11.62
PSAT/NMSQT (Composite)	0.91	0.91 – 0.91
Gender: Female	-8.46	-8.98 – -7.94
Ethnicity: American Indian	-18.43	-22.64 – -14.22
Ethnicity: Asian	4.24	3.38 – 5.10
Ethnicity: Black	-16.26	-17.14 – -15.38
Ethnicity: Hispanic/Latinx	-10.8	-11.49 – -10.11
Ethnicity: Native Hawaiian/Pac. Islander	-17.2	-23.28 – -11.13
Ethnicity: Unknown	-3.16	-5.23 – -1.10
Ethnicity: Two or more	-4.15	-5.40 – -2.89
Parental Education: High School Diploma	2.37	1.24 – 3.50
Parental Education: Associate Degree	5.39	4.31 – 6.47
Parental Education: Bachelor's Degree	12.85	11.78 – 13.92
Parental Education: Graduate Degree	19.35	18.22 – 20.48
Parental Education: No response	-3.11	-5.14 – -1.08
Test Day: School Day	-8.98	-9.53 – -8.44
Weeks since PSAT	0.11	0.09 – 0.13
Leveled up skills	19.57	18.69 – 20.45
Completed practice exam	12.46	11.64 – 13.29
Followed recommended practice	4.44	3.71 – 5.17
Observations	299315	
R_2 / R_2 adjusted	0.865 / 0.865	
AIC	3399192.187	

Appendix E. Modeling the Relationship Between Best Practice Behaviors on OSP and SAT Performance as a Function of Student Characteristics

In this Appendix we conduct a follow-up analysis of best practice behaviors presented in Question 2c of the report. We recommend that the reader reviews that section of the report for context and rationale before reading this Appendix. We also recommend that the reader reviews Appendix D for background on modeling details. Specifically, we examine whether the benefits of engaging in six or more hours of usage with OSP plus at least one best practice behavior extend to all subgroups of students. Thus, we tested whether the best practice spent using OSP usage condition interacted with student characteristics, specifically, gender, ethnicity, parental education, and PSAT/NMSQT. For each interaction, we first fit a base model using only the main effect terms. Then we fit a second model that included the interaction terms. This procedure was repeated for each gender, ethnicity, parental education, and PSAT/NMSQT. As an example, the model specifications for testing the gender X OSP Usage interaction are shown below:

$$(1)\ SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 OSP\ Usage\ Group_i + \varepsilon_i$$

$$(2)\ SAT_i = \beta_0 + \beta_1 PSAT_i + \beta_2 Gender_i + \beta_3 Ethnicity_i + \beta_4 Parental\ Education_i + \beta_5 Test\ Day_i + \beta_6 Weeks\ Since\ PSAT_i + \beta_7 OSP\ Usage\ Group_i + \beta_8 (OSP\ Usage\ Group_i\ x\ Gender_i) + \varepsilon_i$$

There are no terms introduced in these analyses that have not already been described previously, so we refer the reader to Appendix D for further clarification. The interaction coefficient, β_8, represents the moderating effect of demographic variable on OSP Usage Group. Given the number of usage groups and levels of some of the demographic variables, the number of possible interactions are quite large. For parsimony, we are only reporting coefficient estimates for the main effect terms and interaction terms for the 6+ Hour, at least one best practice group. However, other control variables were still included in the models.

The results for the interactions of gender, ethnicity, parental education, and PSAT/NMSQT and OSP hours are shown on Tables E1, E2, E3, and E4, respectively. Visualizations of the marginal means are shown in Figure E1. In general, all interaction models provided an improved fit relative to the main effect models, based on the reduction in AIC. However, we can tell from the R^2Values that these were incredibly modest effects. In none of the interaction models were the changes in R^2even observable when rounding to three digits. When we examine the coefficients of the interaction terms, there are several significant interactions, but all are very small. For example, Table E1 shows that there was reliable interaction between gender and the 6+ Hour, at least one best practice group . However, even if this effect is real, it is so small as to not be practically meaningful. As we see in Figure E1, males saw an estimated benefit of 42 SAT points, whereas females saw a gain of 38 points—a difference of less than 5 points. As with our previous interaction analysis, the fact that the models were able to detect statistically significant effects of these interactions was not surprising given the massive size of this data set. For this reason, we err toward focusing on the practical significance of the results over the statistical significance. To this end, we do not see meaningful differences in benefits derived from the use of OSP across any of the categories examined. One possible exception is from American Indian/Alaska Native students—which saw estimated increases smaller than other racial groups. These students saw increases of around 20 points, whereas other ethnic groups saw increases of 35–41 points. However, as we see in Table E2, the confidence interval for this interaction was very wide, so there is a large degree of uncertainty in the

estimate for this group of students. Unfortunately, as noted in Table 1, there were too few students of this group to provide a reliable estimate of impact.

Table E1.
Regression Results for Gender X Best Practice Group Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Gender: Female	-7.62 ***	-8.01 – -7.23	-6.76 ***	-7.33 – -6.18
OSP Usage: 6+ hrs, w/ at least 1 best-practice behavior	39.22 ***	38.48 – 39.96	41.55 ***	40.42 – 42.69
Female x 6+ Hours, 1+ Best-Practice			-4.01 ***	-5.50 – -2.53
Observations	545640		545640	
R^2 / R^2 adjusted	0.860 / 0.860		0.860 / 0.860	
AIC	6215424.782		6215401.67	
			** p<0.05 ** p<0.01 *** p<0.001*	

Table E2.
Regression Results for Ethnicity X Best Practice Group Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Ethnicity: American Indian	-17.28 ***	-20.44 – -14.13	-14.46 ***	-19.13 – -9.79
Ethnicity: Asian	6.29 ***	5.62 – 6.96	9.03 ***	8.01 – 10.04
Ethnicity: Black	-16.32 ***	-16.99 – -15.65	-14.98 ***	-15.96 – -13.99
Ethnicity: Hispanic/Latinx	-11.00 ***	-11.52 – -10.49	-10.50 ***	-11.22 – -9.78
Ethnicity: Native Hawaiian/Pac. Islander	-13.12 ***	-17.65 – -8.60	-7.11 *	-13.74 – -0.47
Ethnicity: Unknown	-2.49 **	-4.06 – -0.92	-0.65	-2.99 – 1.69
Ethnicity: Two or more	-3.22 ***	-4.17 – -2.27	-1.88 **	-3.31 – -0.45
OSP Usage: 6+ hrs, w/ at least 1 best-practice behavior	39.22 ***	38.48 – 39.96	41.30 ***	40.17 – 42.44
Amer. Ind./Alaska Nat. x <6 hrs. No Best Prac.			0.95	-5.95 – 7.84
Asian x <6 hrs. No Best Prac.			-5.00 ***	-6.53 – -3.48
Black x <6 hrs. No Best Prac.			-1.13	-2.55 – 0.29
Latinx x <6 hrs. No Best Prac.			-0.39	-1.42 – 0.64
Pac. Islander x <6 hrs. No Best Prac.			-9.12	-19.01 – 0.76
Unknown ethnicity X <6 hrs. No Best Prac.			-2.86	-6.34 – 0.62
Two or more races X <6 hrs. No Best Prac.			-3.22 **	-5.35 – -1.09
Amer. Ind./Alaska Nat. x <6 hrs. 1+ Best Prac.			-17.11 **	-28.81 – -5.41
Asian x <6 hrs. 1+ Best Prac.			-5.50 ***	-7.77 – -3.22
Black x <6 hrs. 1+ Best Prac.			-5.99 ***	-8.50 – -3.48
Latinx x <6 hrs. 1+ Best Prac.			-1.22	-2.96 – 0.52
Pac. Islander x <6 hrs. 1+ Best Prac.			-22.34 *	-39.85 – -4.83

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Unknown ethnicity x <6 hrs. 1+ Best Prac.			-5.3	-10.76 – 0.15
Two or more races x <6 hrs. 1+ Best Prac.			-0.94	-4.29 – 2.42
Amer. Ind. x 6+ hrs. No Best Prac.			-22.99 *	-45.69 – -0.30
Asian x 6+ hrs. No Best Prac.			0.44	-4.27 – 5.15
Black x 6+ hrs. No Best Prac.			0.19	-4.35 – 4.74
Latinx x 6+ hrs. No Best Prac.			-2.78	-6.71 – 1.16
Pac. Islander x 6+ hrs. No Best Prac.			-28.46	-66.85 – 9.93
Unknown ethnicity x 6+ hrs. No Best Prac.			-2.2	-13.86 – 9.46
Two or more races x 6+ hrs. No Best Prac.			-2.32	-10.52 – 5.89
Amer. Ind./Alaska Nat. x 6+ hrs. 1+ Best Prac.			-22.97 **	-37.43 – -8.52
Asian x 6+ hrs. 1+ Best Prac.			-4.95 ***	-7.10 – -2.80
Black x 6+ hrs. 1+ Best Prac.			-5.90 ***	-8.34 – -3.46
Latinx x 6+ hrs. 1+ Best Prac.			-2.54 **	-4.45 – -0.63
Pac. Islander x 6+ hrs. 1+ Best Prac.			-5.83	-24.96 – 13.31
Unknown ethnicity x 6+ hrs. 1+ Best Prac.			-3.27	-8.67 – 2.13
Two or more races x 6+ hrs. 1+ Best Prac.			-0.74	-4.31 – 2.83
Observations	545640		545640	
R^2 / R^2 adjusted	0.860 / 0.860		0.860 / 0.860	
AIC	6215424.782		6215348.669	

** p<0.05 ** p<0.01 *** p<0.001*

Table E3.
Regression Results for Parental Education X Best Practice Group Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Parental Education: High School Diploma	2.47 ***	1.63 – 3.32	2.24 ***	1.01 – 3.48
Parental Education: Associate Degree	5.39 ***	4.59 – 6.20	4.89 ***	3.72 – 6.06
Parental Education: Bachelor's Degree	13.68 ***	12.87 – 14.48	13.45 ***	12.32 – 14.58
Parental Education: Graduate Degree	20.84 ***	19.99 – 21.69	20.76 ***	19.58 – 21.95
Parental Education: No response	-1.62 *	-3.12 – -0.12	0.05	-2.10 – 2.21
OSP Usage: 6+ hrs, w/ at least 1 best-practice behavior	39.22 ***	38.48 – 39.96	38.89 ***	36.16 – 41.63
High School Diploma x 6+ hours, w/ at least 1 best-practice			-0.45	-3.95 – 3.04
Associate Deg. x 6+ hours, w/ at least 1 best-practice			-0.59	-3.84 – 2.65
Bachelor's Deg. x 6+ hours, w/ at least 1 best-practice			1.4	-1.62 – 4.42
Graduate Deg. x 6+ hours, w/ at least 1 best-practice			0.36	-2.68 – 3.40
No response x 6+ hours, w/ at least 1 best-practice			-5.34	-11.39 – 0.71
Observations	545640		545640	
R^2 / R^2 adjusted	0.860 / 0.860		0.860 / 0.860	
AIC	6215424.782		6215437.77	
			** p<0.05 ** p<0.01 *** p<0.001*	

Table E4.
Regression Results for PSAT/NMSQT X Best Practice Group Interactions

	Main Effect		*Interactions*	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
PSAT/NMSQT (Composite)	0.91 ***	0.91 – 0.91	0.90 ***	0.90 – 0.90
OSP Usage: 6+ hrs, w/ at least 1 best-practice behavior	39.22 ***	38.48 – 39.96	37.96 ***	37.18 – 38.73
PSAT/NMSQT x 6+ hours, w/ at least 1 best-practice behavior			0.03 ***	0.02 – 0.03
Observations	545640		545640	
R^2 / R^2 adjusted	0.860 / 0.860		0.860 / 0.860	
AIC	6215424.782		6215119.135	
			** p<0.05 ** p<0.01 *** p<0.001*	

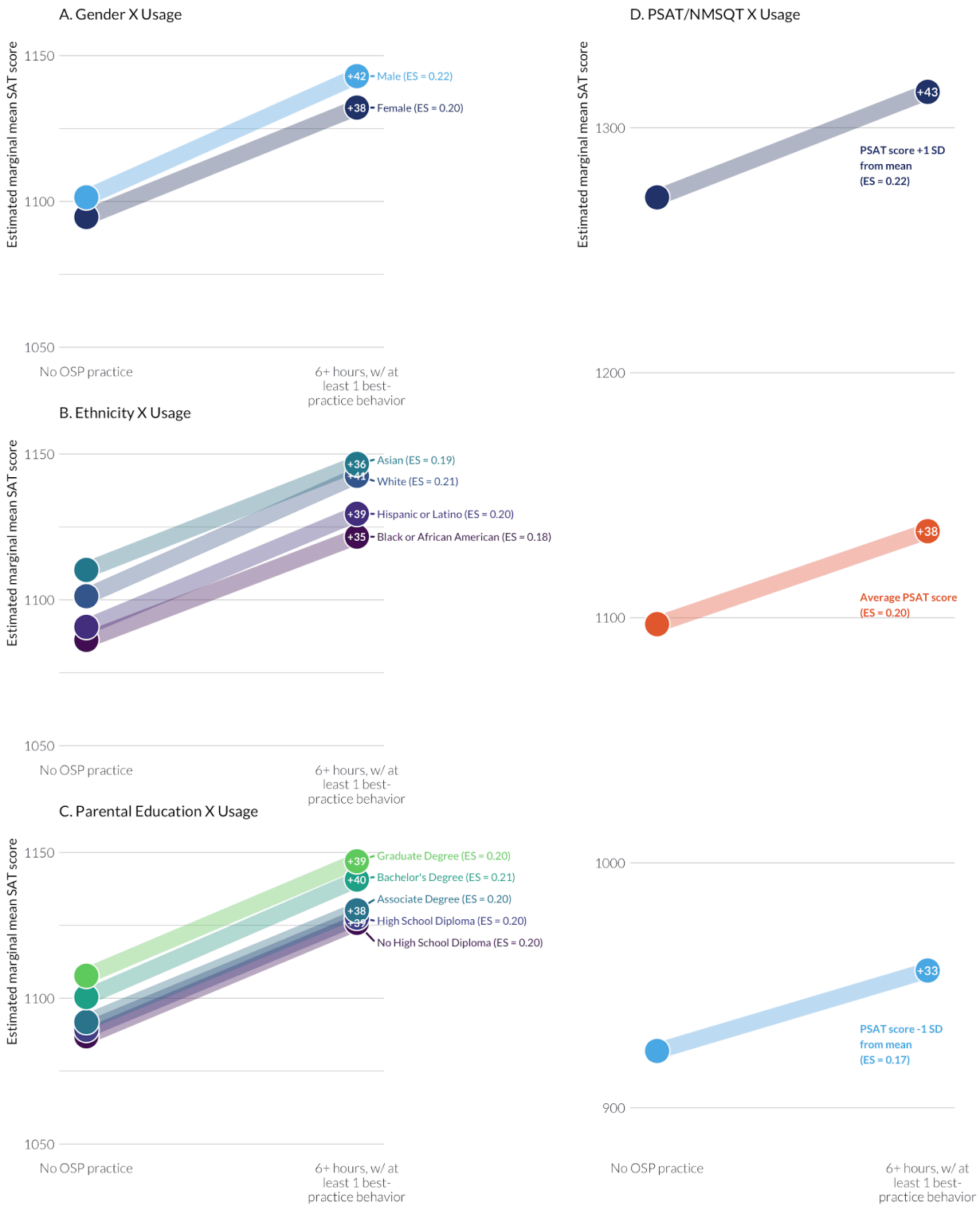


Figure E1. Visualizations of best practice conditions by student characteristics interactions. Note that only the largest subgroups are shown.

Appendix F. Sensitivity Analysis

In this report, we have explored data from students who have taken the PSAT/NMSQT and the SAT, and have engaged in OSP usage in a real-world setting. This was by definition an observational study, as students were free to choose their own engagement with the OSP platform.

In contrast to an observational study, a randomized controlled trial might assign students to different levels of OSP usage and different combinations of best practice behaviors. Such a research design would allow us to test the impact of OSP usage while using random assignment to ensure that hidden confounds are not systematically represented in, or responsible for assignment to, the different experimental groups. However, such a design was neither practically nor ethically possible for this study. Additionally, observational and quasi-experimental studies are often more powerful for the generalizability of findings while providing weaker causal evidence. Working with observational data allows us to directly observe the real-world behavior of a large number of students, rather than relying on artificially constructed, assigned behavior. However, without the ability to randomly assign students to different OSP usage conditions and best practice behaviors, it remains possible that systematic differences between usage groups could underlie both their levels of OSP usage and SAT achievement.

This conflict between natural behavior and causal inference is a key tension in observational research. Although our inclusion of PSAT/NMSQT scores does allow us to include a strong measure that helps control for prior academic achievement in our models, there is still a challenge in evaluating the impact of OSP when we are unable to assign students to the OSP usage groups. As a step toward addressing this challenge, this Appendix presents a propensity score analysis that supplements our key OSP research question: *Is usage of OSP related to improved SAT performance?*

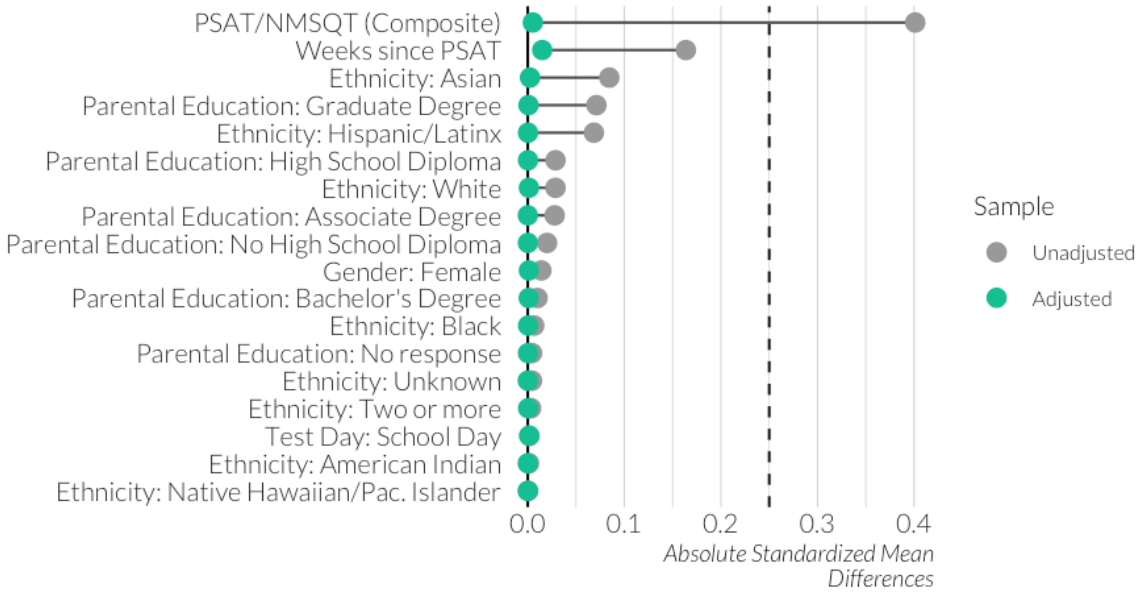
Propensity score matching or weighting is a method that allows for the estimation of causal effects in observational data by creating a propensity score variable that predicts the treatment status by accounting for the covariates in the regression model. This new variable is used to generate a balanced representation of covariates between the treatment and control conditions. There are various techniques that address this question. The method in this Appendix can be understood as the one attempting to account for overt sampling bias in quasi-experimental designs, which estimates the conditional probability of receiving treatment based on the measured covariates in our sample (Rosenbaum and Rubin, 1983; Guo & Fraser, 2015; Thoemmes & Ong, 2016). Note that propensity scores are an estimation; this variable is not a measured attribute within the data set. Rather it is a calculation that predicts the probability of being in the treatment versus comparison condition for the participants in the study. In effect, these methods scrutinize our previous estimate of the treatment effect (in this case, effective OSP usage) with a higher bar, particularly asking whether adjusting for covariate imbalances might change the estimated increase in SAT outcomes associated with effective OSP usage.

Design

For the propensity score analysis, we subselected students who completed at least six hours and one best practice on the platform, and then compared them to students with no OSP usage. This is because we wanted to target the effect of OSP usage *under ideal conditions* on SAT score improvement, and to learn whether the estimates of that OSP usage would change, given balanced covariates. As in the previous models presented in the main body of this report and in subsequent discussion of usage groups, this student group that had completed at least “six hours and one best practice” can also be imagined as the

group that received the strongest “dose” of our intervention. In order to best approximate a comparison to a controlled experiment design, we chose this to represent a complete treatment of the intervention.

Figure F. The covariate balance before and after the application of propensity score weights. Of the covariates, only PSAT/NMSQT passes the threshold for severe imbalance.



Analysis

We conducted multiple propensity score models using both logistic regression and gbm weighting methods, with results shown below (Tables F1 and F2). **In summary, our original treatment effect estimates were not dramatically changed by the additional scrutiny of propensity weights regardless of the method:** our unweighted model estimate for the treatment effect of six hours plus one best practice was an additional increase in SAT points of approximately **39,** and our estimates below range from **35–39.** Inverse probability weights were generated for both logistic regression and gbm models using the weightit R package (Greifer, 2020). The estimate of confidence interval was generated using the 'robust' method (Robins et al., 2000; Hainmueller, 2012) via the survey R package (Lumley, 2020). It is important to note that this analysis is being used as a sensitivity check on our main treatment effect estimates within the context of an observational design, and does not provide a substitute for the random assignment of participants to treatment conditions. That said, the results indicate that the treatment effects hold when using statistical adjustment to account for covariate differences in the likelihood for whether someone will use the OSP platform as intended.

Table F1.
Treatment Effect Estimated by Propensity Score Method ATT

Model	Estimate	Covariates	ESS	Confidence Interval
Logistic Regression	37.6	None	Weighted: 174,845	35.6 - 39.7
	38.6	All usual predictors	Unweighted: 246,325	37.9 - 39.3
GBM	35.7	None	Weighted: 150,435	35.7 - 37.7
	35.8	All usual predictors	Unweighted: 246,325	35.8 - 36.6

Table F2.
Treatment Effects Estimates by Propensity Score Method ATE

Model	Estimate	Covariates	ESS	Confidence Interval
Logistic Regression	36.7	None	Weighted: 35,681	34.5 - 39.0
	37.5	All usual predictors	Unweighted: 43,946	36.6 - 38.3

Appendix G. Modeling the Relationships Between Student Characteristics and the Likelihood of Engaging in Best Practice Behaviors

In this Appendix, we present details for the follow-up analysis of the likelihood of engaging in either six or more hours of OSP usage, or the three best practice behaviors presented in subsection 2d of the report. In this analysis, we defined each outcome measure in categorical terms, as either present or not present for linkers who finished at least one problem on OSP. Across four logistic regression models, we examined the odds ratio for each predictor (e.g., gender, ethnicity, parental education, test administration) on the likelihood of completing either the time measure or the best practice. These odds ratios are represented in Figure 12 in the referenced subsection in the main body of the report, and are also reported below in the more comprehensive Table G.

Table G.
Odds ratios for each Predictor Variable Across Outcome Measures of Time Spent on OSP, and Best Practice Behaviors.

	OSP Usage: 6+ hrs		*15+ Skills Leveled Up*		*Completed Practice Exam*		*10+ Tasks Majority Recommended*	
Predictors	*Odds Ratios*	*CI*	*Odds Ratios*	*CI*	*Odds Ratios*	*CI*	*Odds Ratios*	*CI*
Intercept	0.1	0.1 – 0.1	0.0	0.0 – 0.0	0.1	0.1 – 0.1	0.1	0.1 – 0.1
PSAT/NMSQT (Composite)	1.3	1.3 – 1.3	1.9	1.9 – 2.0	1.3	1.2 – 1.3	1.2	1.2 – 1.2
Gender: Female	1.0	1.0 – 1.0	1.2	1.2 – 1.2	1.0	1.0 – 1.1	1.0	1.0 – 1.0
Ethnicity: American Indian	1.1	0.9 – 1.3	0.6	0.4 – 0.8	1.1	0.9 – 1.3	1.0	0.8 – 1.2
Ethnicity: Asian	2.0	2.0 – 2.1	0.7	0.6 – 0.7	0.8	0.7 – 0.8	0.9	0.8 – 0.9
Ethnicity: Black	1.7	1.6 – 1.7	0.6	0.6 – 0.6	0.9	0.8 – 0.9	0.8	0.8 – 0.8
Ethnicity: Hispanic/Latinx	1.2	1.2 – 1.3	0.8	0.7 – 0.8	0.8	0.8 – 0.9	0.9	0.9 – 0.9
Ethnicity: Native Hawaiian/Pac. Islander	1.1	0.9 – 1.4	0.5	0.4 – 0.8	0.7	0.5 – 1.0	1.1	0.8 – 1.4
Ethnicity: Unknown	1.6	1.4 – 1.7	0.8	0.7 – 0.8	0.8	0.8 – 0.9	1.0	0.9 – 1.1
Ethnicity: Two or more	1.2	1.1 – 1.2	0.8	0.8 – 0.9	0.9	0.9 – 1.0	1.0	1.0 – 1.1
Parental Education: High School Diploma	0.9	0.9 – 1.0	1.1	1.1 – 1.2	1.0	1.0 – 1.1	1.0	1.0 – 1.0
Parental Education: Associate Degree	0.9	0.9 – 1.0	1.1	1.1 – 1.2	1.0	1.0 – 1.1	1.0	1.0 – 1.1
Parental Education: Bachelor's Degree	1.0	1.0 – 1.1	1.2	1.1 – 1.3	1.1	1.0 – 1.1	1.1	1.1 – 1.2
Parental Education: Graduate Degree	1.2	1.2 – 1.3	1.2	1.1 – 1.3	1.1	1.0 – 1.1	1.1	1.1 – 1.2
Parental Education: No response	1.0	0.9 – 1.1	1.2	1.0 – 1.3	1.0	0.9 – 1.1	1.1	1.0 – 1.2
Test Day: School Day	1.1	1.1 – 1.2	1.5	1.4 – 1.5	0.9	0.9 – 1.0	1.0	1.0 – 1.0
Weeks since PSAT	1.0	1.0 – 1.0	1.0	1.0 – 1.0	1.0	1.0 – 1.0	1.0	1.0 – 1.0
OSP hours			1.3	1.3 – 1.3	1.2	1.2 – 1.2	1.2	1.2 – 1.2

	OSP Usage: 6+ hrs	*15+ Skills Leveled Up*	*Completed Practice Exam*	*10+ Tasks Majority Recommended*
Observations	299315	299315	299315	299315
AIC	273341.347	161519.781	190251.308	240518.903