CollegeBoard
connect to college success™

Research Notes
Office of Research and Analysis RN-33, ETS RM-07-02, August 2007

Effects of Extra Time on Performance on New SAT® Questions

Brent Bridgeman and Frederick Cline

Introduction

The SAT® is intended to be a power test in which speed of responding plays at most a minor role in determining test scores (Donlon, 1984). Because questions that are unanswered because of time difficulties (i.e., test speededness) are of no use in estimating a student's abilities (except to the extent that speed itself is a relevant ability), efforts are made to assure that the test is not too speeded. The recent decision by the College Board to no longer flag scores of disabled examinees who are granted extra time (typically time and a half) further reinforces the notion that speed of performance is not seen as a key component of the reasoning abilities assessed by the SAT. Although guidelines that focus on completion rates, such as those proposed by Swineford (1974), have proven to be useful in identifying highly speeded tests, meeting the guidelines does not assure that speed is a trivial component of the scores. Answering the last question does not guarantee that all items were fully considered, as students are advised to skip items that appear to be too time consuming. Students who followed this advice and answered the last item just as time ran out by skipping earlier items would be counted as finishing the test by the Swineford guidelines, though they might have gotten higher scores if they had time to revisit the skipped questions. Even if nearly all students answer every question, scores could still be affected by rushing at the end.

The current study took an experimental approach to evaluating test speededness. In order to assess the benefits of extra time (or the penalty of strict time limits) on new SAT scores, sections that were designed to be administered with a 25-minute time limit were administered with a 40-minute time limit (or slightly more than time and a half) as part of the SAT field trial that was administered to thousands of eleventh-grade students in the spring of 2003.

Method

Sample and Design

Details of the field trial design are provided elsewhere (Liu et al., 2002). The current analysis used Design 2. There were 10 booklets in this design. Although total testing time for each booklet was the same (215 minutes), the timing of tests within sets of booklets varied, requiring participating schools to provide three separate rooms to accommodate the three timing requirements. Schools were asked to randomly assign students to rooms. It appears that this random assignment was accomplished successfully, as differences on common sections across rooms were very small (Liu and Feigenbaum, 2003). Four test sections (one in each of four booklets) were evaluated with extended time: New Reading (NR1), which included both reading item types for the new test (reading passages and sentence completions); two sections of New Mathematics (NM1, which contained all five-choice multiple-choice [MC] questions, and NM2, which included both MC items and items in which students produce a numerical response [SPR]); and the multiple-choice section of the New Writing section (NW[MC]). Standard-time versions of these four sections were included in other booklets; sample sizes were larger for the standard-time sections because a given section could appear in more than one booklet, and sample sizes were larger for some booklets than others.

Data Cleaning

When a low-stakes test (e.g., one in which scores are not reported to colleges) is administered, there is a concern that scores may be invalid because the study participants are not making a serious effort. For the field trial this did not appear to be a major problem because the participants were

high school juniors who would view the test as an opportunity to prepare for the SAT under realistic testing conditions with items that are generally of the same type that they would encounter on the real test. We screened out the few students who apparently were not taking the test seriously because they attempted to answer only a few questions in a section. As a check on the validity of the scores from the field test, correlations between scores on the field trial test and the full operational PSAT/NMSQT® (P/N) that most participants[1] had taken about six months previously were obtained. Correlations between P/N sections and scores on the short field trial sections were quite high, suggesting that the field trial test assessed reading, math, and writing abilities in a manner that was fully consistent with the measurement on the operational test. Furthermore, as indicated in Table 1, correlations were comparable in the standard- and extended-time sections.

Analyses

Differences in formula scores across timing conditions were evaluated with analysis of variances (ANOVAs). In addition to timing condition, independent variables included gender, race/ethnicity (African American, Asian American, Hispanic, white, and other), and ability as estimated from the comparable P/N test. Three levels of ability were defined for each of the three scores on the P/N. The levels were defined such that half of the students would be in the middle group, with a quarter in each of the two extreme groups. Score ranges in each level were as follows: Level 1 = 20–42; Level 2 = 43–56; Level 3 = 57–80. (Identical ranges were used for all three P/N scores.) Sample sizes by subgroup for NR1 are summarized in Table 2. By design, sample sizes for the other three test sections are similar.

Results and Discussion

New Reading (NR1) Results

Extra time provided absolutely no advantage on the 24-item New Reading section (NR1). Indeed, the mean formula score was actually slightly higher in the standard-time group than in the extended-time group (see Table 3). Differences (or more accurately, nondifferences) were comparable across gender and racial/ethnic groups, and the ANOVA indicated no significant interactions of gender or race/ethnicity with the timing condi-

[1] P/N scores were available for about 85 percent of the examinees tested. Less than 10 percent of these had taken the P/N in 2001; the remainder were tested in October of 2002.

Table 1

Correlations of Experimental Section Scores with Operational Scores on Comparable Section of PSAT/NMSQT for Standard- and Extended-Time Conditions

	Standard Time		*Extended Time*	
Tests Correlated	**n**	**r**	**n**	**r**
NR1 and P/N-V*	2,562	.80	614	.77
NM1 and P/N-M**	2,476	.77	675	.80
NM2 and P/N-M	2,476	.77	656	.78
NW(MC) and P/N-W***	2,401	.79	585	.77

*PSAT/NMSQT verbal section. **PSAT/NMSQT mathematics section
*** PSAT/NMSQT writing section.

Table 2

Sample Sizes by Subgroup for NR1

P/N Level	*Race/Ethnicity*	*Gender*	*Standard Time* n	*Extended Time* n
1	African American	M	61	19
		F	139	23
	Asian American	M	20	5
		F	31	8
	Hispanic	M	58	11
		F	74	26
	White	M	80	28
		F	134	40
	Other	M	21	3
		F	27	5
2	African American	M	58	14
		F	105	26
	Asian American	M	36	12
		F	50	9
	Hispanic	M	94	27
		F	100	24
	White	M	290	76
		F	424	110
	Other	M	34	5
		F	46	7
3	African American	M	14	2
		F	12	2
	Asian American	M	25	7
		F	29	4
	Hispanic	M	30	9
		F	34	3
	White	M	186	47
		F	242	39
	Other	M	21	3
		F	31	9
Total			**2,506**	**603**

Table 3

Mean (SD) Formula Scores for Standard-Time and Extended-Time Sections

Test	*Standard Time*	*Extended Time*	*Formula Score Difference*	*Difference in SD Units (d)*
NR1	10.63 (6.2)	10.46 (6.0)	-0.17	-0.03
NM1	9.20 (5.0)	9.65 (5.1)	0.45	0.09
NM2	7.01 (4.6)	7.40 (4.7)	0.39	0.08
NW(MC)	13.48 (8.5)	14.85 (8.9)	1.37	0.16

tions. The P/N level did not interact significantly with the time condition (F [2, 3170] = 2.00, p = .135). For the last couple of items, the proportion correct was only slightly higher in the extended-time forms (.36 to .41 on the last item, and .51 to .55 on the penultimate item), and the proportion incorrect (as opposed to omitted) was also slightly higher in the extended-time form, accounting for the slight formula score advantage in the standard-time group. The very small differences on the New Reading section are consistent with the very small differences associated with extra time on the current SAT critical reading section (Bridgeman, Trapani, and Curley, 2004).

New Mathematics Results (NM1 and NM2)

NM1 contained only MC items. Scores were higher in the extended-time condition by less than half of a formula score point. Similarly, scores for NM2 (which includes SPR items) were less than half a formula score point higher in the extended-time condition. These differences were slightly smaller than differences noted in the current SAT mathematics section (Bridgeman, Trapani, and Curley, 2004), suggesting that efforts to create a less speeded mathematics section were successful. ANOVA results indicated that the difference between timing conditions was statistically significant for both NM1 (F [1, 3,045] = 9.36, p < .01) and NM2 (F [1, 3,029] = 12.40, p <.01). Interactions of timing condition with gender, race/ethnicity, and ability levels were not significant (ps > .05). Proportion correct on the last four items (all SPRs) were slightly higher for the examinees in the extended-time group: .26 versus .33; .21 versus .25; .15 versus .20; and .06 versus .10.

New Writing (Multiple Choice) (NW[MC]) Results

Extra time provided a noticeable advantage on the writing scores with scores about 1.4 formula score points higher in the extended-time group, though even this difference of 0.16 standard deviation units is small by conventional standards. In the ANOVA, the main effect of timing condition was significant (F [1, 2,877] = 7.95, p = .005), as was the interaction with ability level (F [2, 2877] = 3.75, p = .024), but interactions with gender and race/ethnicity were not significant. Table 4 shows the mean scores in both timing conditions by ability level on the P/N. Standardized differences (d) use the total group (rather than within ability group) standard deviation to better reflect the effect of extra time on total scores. Thus, the difference of .25 SD units in the middle-ability group would translate to a difference of about 28 points on a test with an SD of 110 (such as the SAT). Differences were trivial in the high- and low-ability groups. These results should be interpreted cautiously because the P/N itself could be speeded; to the extent that high scores on the P/N require speed as well as recognition of the conventions of Standard Written English, the high-ability group could also be a fast group in which extra time would not be expected to make much of a difference.

Table 5 shows the proportion of students selecting the correct answer, incorrect answer, or no answer (omitted or not reached) for the last 20 items in the section for students in both timing conditions. From item 27 through the end of the test, the proportion correct is at least .05 points higher for the participants with extended time. On three items, the difference is .10 or greater. Given these relatively large differences, it may be somewhat surprising that the difference in formula scores is only 1.37 points. The explanation may lie in the "% Incorrect" column, as students with extra time are also considerably more likely to get the items toward the end of the test wrong (as opposed to omitted) when they have more time. From item 28 on, there are at least three times as many students attempting to answer each item in the group with extended-time limits as in the group with standard-time limits. The percent of students not attempting an item in the standard- and extended-time groups can be seen graphically in Figure 1.

Table 4

Mean (SD) Formula Scores for Standard-Time and Extended-Time Sections for the New Multiple-Choice Writing Section by Score Level on the P/N-W

P/N-W Level	*Standard Time*	*Extended Time*	*Formula Score Difference*	*Difference in Total Group SD Units (d)*
1 (20–42)	6.67 (5.0)	7.05 (5.5)	0.38	0.04
2 (43–56)	14.04 (6.3)	16.20 (6.4)	2.16	0.25
3 (57–80)	24.16 (5.5)	24.64 (6.3)	0.48	0.06

Table 5

Proportion Correct, Incorrect, and Not Answered for the Last 20 New Writing Multiple-Choice Items

	% Correct		% Incorrect		% No Answer	
Item	*Standard*	*Extended*	*Standard*	*Extended*	*Standard*	*Extended*
18	.56	.56	.37	.39	.08	.05
19	.69	.70	.25	.28	.06	.03
20	.73	.72	.22	.26	.06	.02
21	.41	.43	.53	.54	.07	.03
22	.33	.36	.59	.61	.08	.03
23	.27	.28	.65	.69	.08	.03
24	.33	.38	.57	.59	.10	.04
25	.32	.36	.56	.60	.13	.04
26	.64	.66	.23	.30	.13	.05
27	.45	.50	.42	.45	.14	.05
28	.42	.48	.44	.48	.14	.04
29	.69	.78	.14	.18	.17	.04
30	.65	.73	.16	.21	.19	.06
31	.25	.30	.53	.66	.22	.05
32	.38	.43	.31	.48	.31	.09
33	.37	.49	.33	.44	.31	.07
34	.49	.69	.17	.23	.34	.07
35	.30	.37	.33	.54	.38	.09
36	.36	.55	.23	.35	.41	.10
37	.28	.37	.28	.51	.44	.12

Conclusion

As estimated from the improvement in scores with extra time, time limits for the New Reading section are appropriate in that additional time does not impact performance. Performance on both parts of the New Mathematics section appears to be only slightly enhanced with additional time. The proportion correct on the last four SPR items (which are not influenced by increased guessing with extra time) improved by .04 to .07 points, suggesting that a small time extension would help some students to more fully demonstrate their abilities. Nevertheless, scores increased by less than half of a formula score point with an extra 15 minutes of testing time; by this standard, the current time limit could be considered adequate. With the existing time limits, test users could be assured that extra time provides at best a very modest advantage for nondisabled test-takers on the New Reading and Mathematics sections. The largest differences were noted on the New Writing section. Formula scores improved about 1.37 points with extra time. This analysis confirms the conclusions of Allspach and Walker (2003). Following a very detailed analysis of conventional speededness indices, as well as an analysis indicating a higher level of performance on a 30-minute section than on a 25-minute section, they concluded

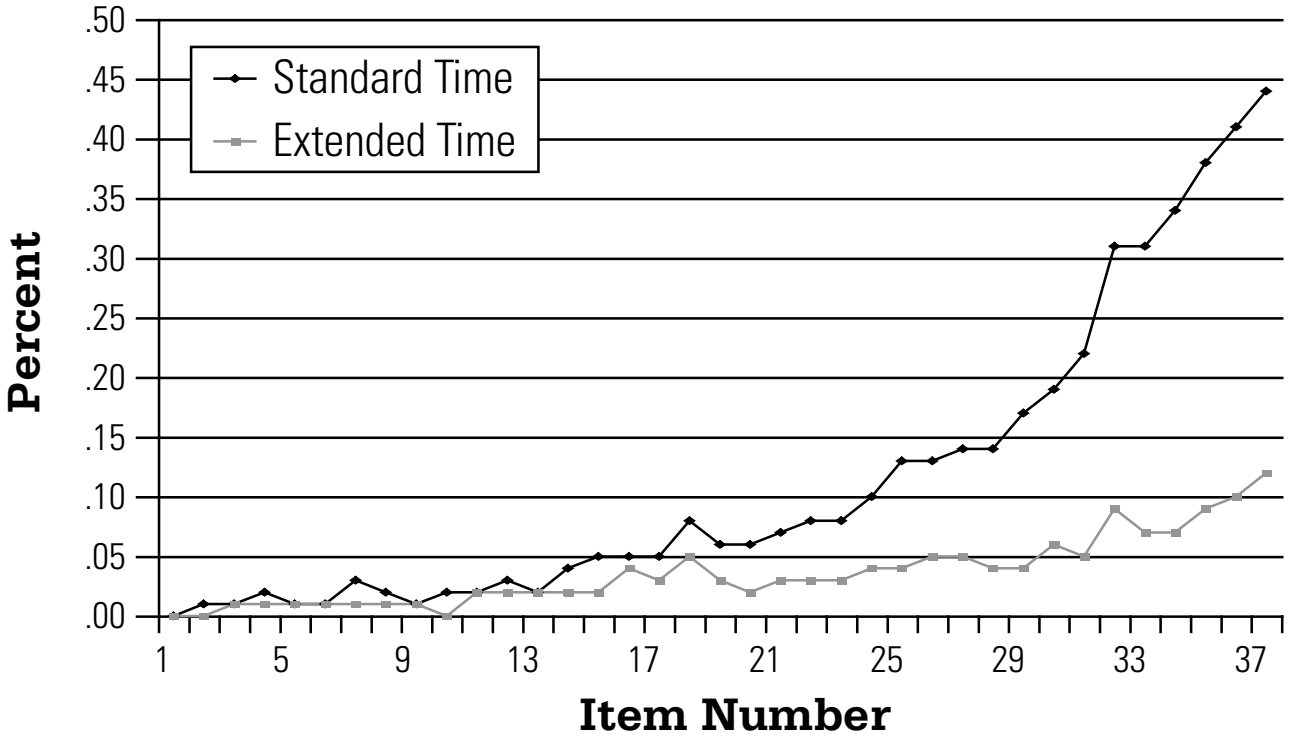

Figure 1. Percent of students not answering in standard-time and extended-time groups on the New Writing section.

that the 25-minute test was speeded and that "students would benefit from increased time on the test" (p. 28). As they suggest, extra time might be provided by reducing the number of items on the test or by using less time-consuming item formats.

References

Allspach, J. R. & Walker, M. E. (2003). *Assessing speededness of the writing section of the new SAT* (SR-2003-69). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41* (4), 291–310.

Donlon, T. F. (Ed.) (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: The College Board.

Liu, J., Feigenbaum, M., Walker, M., Baron, P., Cook, L., & Wendler, C. (2002, September). *Proposal for the first field trial design for SAT and PSAT/NMSQT*. Unpublished memorandum. Princeton, NJ: Educational Testing Service.

Liu, J. & Feigenbaum, M. (2003). *Prototype analysis of spring 2003 New SAT field trial* (ETS Statistical Report SR-2003-69). Princeton, NJ: Educational Testing Service.

Swineford, F. (1974). *The test analysis manual* (ETS Statistical Report SR-74-06). Princeton, NJ: Educational Testing Service.

Office of Research and Analysis
The College Board
45 Columbus Avenue
New York, NY 10023-6992
212 713-8000

The College Board: Connecting Students to College Success
The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,200 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns. For further information, visit www.collegeboard.com.