The College Board

Research Report No. 2003-2

Effect of Fewer Questions per Section on SAT® I Scores

Brent Bridgeman, Catherine Trapani, and Edward Curley

College Board Research Report No. 2003-2
ETS RR-03-08

Effect of Fewer Questions per Section on SAT® I Scores

Brent Bridgeman, Catherine Trapani, and Edward Curley

College Entrance Examination Board, New York, 2003

Brent Bridgeman is principal research scientist at Educational Testing Service (ETS).

Catherine Trapani is associate research data analyst at ETS.

Edward Curley is director–verbal skills at ETS.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board: Expanding College Opportunity

The College Board is a national nonprofit membership association whose mission is to prepare, inspire, and connect students to college and opportunity. Founded in 1900, the association is composed of more than 4,300 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three 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.

Additional copies of this report (item #996735) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Contents

Tables

Figures

Abstract

The impact of allowing more time for each question on SAT® I: Reasoning Test scores was estimated by embedding sections with a reduced number of questions into the standard 30-minute equating section of two national test administrations. Thus, for example, questions were deleted from a verbal section that contained 35 questions to produce forms that contained 27 or 23 questions. Scores on the 23-question section could then be compared to scores on the same 23 questions when they were embedded in a section that contained 27 or 35 questions. Similarly, questions were deleted from a 25-question math section to form sections of 20 and 17 questions. Allowing more time per question had a minimal impact on verbal scores, producing gains of less than 10 points on the 200–800 SAT scale. Gains for the math score were less than 30 points. High-scoring students tended to benefit more than lower-scoring students, with extra time creating no increase in scores for students with SAT scores of 400 or lower. Ethnic/racial and gender differences were neither increased nor reduced with extra time.

Effect of Extra Time on SAT® I Scores

The SAT I: Reasoning Test (SAT) assesses verbal and mathematical reasoning skills that are predictive of success in college. According to the technical handbook for the SAT, the speed with which students can answer the questions should play at most a minor role in determining scores (Donlon, 1984). Although time limits could affect the scores of all students, the possibility of differential effects for females and minority students has been a particular concern. For example, it has been suggested that females may be at a disadvantage on the mathematical portion of the SAT because they use time-intensive algorithmic strategies (Linn, 1992) or allocate their time inefficiently (Becker, 1990). Thus, in addition to knowing the general impact of allowing extra time, the differential impact for ethnic and gender groups is also of interest.

There is a common belief that if examinees had only a little more time they could substantially improve their scores. The number of SAT examinees requesting extra time (which is provided to students with documented disabilities who require additional testing time) has grown by about 26 percent over the past five years. Though this increase is to be expected as more students with disabilities consider college as an option, there is still a concern that it provides an opportunity for abuse from those seeking to improve scores by any means possible. It is extremely difficult, if not impossible, to clearly separate the students with legitimate disabilities from those who are gaming the system. However, the issue would be moot if extra time had little impact on test scores. If there were credible evidence that extra time does not affect student performance, there would be little or no motivation to manipulate the system to gain extra time.

Methods for determining the impact of time limits on test scores, often referred to as speededness, rely either on completion data from a single administration or on an experimental manipulation of testing time. The guidelines used routinely for evaluating speededness on the SAT (Swineford, 1974) are of the former type, and specify that in order to be considered unspeeded virtually all of the students should respond to at least one question beyond three-fourths of the way through a section, and at least 80 percent of the students should respond to the last question. Although these guidelines can be useful for identifying very speeded test forms, meeting the guidelines does not assure that speed is a trivial component of the scores. Whether using the Swineford guidelines or other non-experimental approaches (Rindler, 1979), certain assumptions are required that are unlikely to be fully met in practice. One critical assumption is that the questions are answered in the order presented. However, suppose a student skips items that appear to be time consuming, intending to return to them at the end of the test, but that time runs out just as the student is answering the last question. Because the student answered the last question, the internal criteria suggest that the test is unspeeded for such students, though they might get higher scores if they had time to revisit the skipped questions. Even if no questions are skipped, scores might still be substantially different if the examinee had time to consider each question more fully.

A quasi-experimental approach was used to determine how much students classified as learning disabled gained when they initially took a test with regular timing and then took an extended-time test (Camara, Copeland, and Rothschild, 1998). These students made greater gains than are typical for students who merely repeat the test with the same timing conditions each time, but the effects of self selection on a sample that chooses to take the test once with regular timing and then requests extra time is unknown. Also, this study was limited to students with a disability classification and could not estimate whether comparable effects would be found for nondisabled students.

True experimental studies permit a direct evaluation of the impact of extra time, but they are difficult to carry out under realistic testing conditions. One study that manipulated time on SAT questions found that providing an extra 10 minutes on 30-minute math and verbal sections did not produce a statistically significant benefit (Evans, 1980), but sample size was limited to only 36 students per section resulting in little power to detect small differences. A much larger scale study of verbal and math scores on the Graduate Record Examination General Test found that an extra 10 minutes on a 20-minute section increased scores by less than one point each on both the 26-question verbal test and the 14-question quantitative test (Wild, Durso, and Rubin, 1982). Extended time did not interact with either gender or race (black/white). A recent research summary suggests that there is no evidence that extending time limits benefits minority subgroups, but that there is some evidence that extending time limits is sometimes detrimental to minority subgroups (Sackett, Schmitt, Ellingston, and Kabin, 2001). However, their review did not identify any research on this issue for high-stakes admissions tests during the last 20 years.

Effects of extra time may be studied experimentally either by administering the same number questions with additional time, or by keeping the time constant but reducing the number of questions. The studies reported here used the latter experimental approach by embedding sections with a reduced number of questions into the 30-minute equating section of national administrations of the SAT. This section is used for a variety of purposes, such as test equating and evaluating the psychometric characteristics of new questions. Although this section does not contribute to the reported scores, examinees are not told which section is the equating section, so they are fully motivated to do their best. Data were obtained from tests administered in the fall of 2000 (Study 1), and in a follow-up study in the fall of 2001 (Study 2) that used exactly the same procedures but a different set of questions,

Method

Test Forms

Every operational form of the SAT test includes two 30-minute verbal sections. These sections (V1 and V2) both contain the same question types (analogies, sentence completions, and critical reading), though V1 contains more questions and has a lower proportion of questions based on reading passages. Similarly, every operational form of the SAT includes two 30-minute math sections (M1 and M2). M1 contains 25 five-choice questions, and M2 contains 15 four-choice quantitative comparison (QC) questions and 10 questions with a student-produced response (SPR) in which the examinee grids a numerical value on the answer sheet rather than making a multiple-choice selection. Each administration of the SAT also contains a 15-minute verbal section and a 15-minute math section plus a 30-minute equating section that contains verbal questions for some examinees and math questions for other examinees. Test booklets are packaged so that different versions (or spirals) of the equating section can be essentially randomly distributed. Ten spirals in the fall of 2000 and 10 spirals in the fall of 2001 were used to address the speededness issue.

For the purposes of these studies, shortened test forms were created from previously administered test forms by deleting questions at different difficulty levels so that the difficulty levels and range of difficulty (mean and SD of the equated deltas) of the original and shortened versions were essentially the same.

The 10 forms administered in each of these studies were as follows:

1. V1 standard 35-item length
2. V2 standard 30-item length
3. M1 standard 25-item length
4. M2 standard 25-item length (15 QC and 10 SPR)
5. V1 shortened to 27 items
6. V2 shortened to 25 items
7. M1 shortened to 20 items
8. M2 shortened to 22 items (because of the design of the answer sheet, we administered all 15 QC items; 3 SPRs [which are relatively time-consuming] were deleted)
9. V1 shortened to 23 items
10. M1 shortened to 17 items

The order of the common items was the same in all forms (e.g., common item 1 was always administered before common item 2), though the actual item numbers were necessarily different in the different forms (e.g., the item in the 10 position in the shortest form was in the 15 position in the longest form). However, forms were designed so that the last item was identical in the original and shortened forms (e.g., item 35 in form 1 was the same as item 27 in form 5 and item 23 in form 9).

The level of speededness reduction reflected in forms 5–8 is a realistic level of reduction for a future operational test that could be administered in the same time as the current test and with adequate reliability. (A test composed of the shortened sections would have two more items than the current PSAT/NMSQT® even before items from the two 15-minute sections were added on, thus virtually assuring a higher level of reliability than the current PSAT/NMSQT. The PSAT/NMSQT contains the same item types as SAT I and is currently used as a practice test for the SAT I and as a preliminary screen for merit scholarships.)

The level of reduction in forms 9 and 10 is approximately equivalent to allowing time-and-a-half for the current test. This level of reduction may be problematic for the design of an operational test, but it provides crucially important information for client institutions that are concerned with how to interpret scores on extended time tests.

Data Source

Each spiral in Study 1 contained at least 8,000 examinees and the Study 2 spirals were slightly larger. Because examinees were randomly assigned to spirals, the ethnic and gender composition of each of the spirals was comparable; therefore, to simplify the presentation, Table 1 contains ethnic by gender sample sizes only for the first verbal spiral of Study 1. For analysis purposes, examinees were divided into three ability groups by their scaled scores on the operational verbal sections. The groups were: less than 410, 410–600, and greater than 600. These divisions are also reflected in Table 1. Although the bottom category covers more score points, there are more people in the top category because of a negative skew in the overall population and because the test administration selected for the study attracted a disproportionate representation of high-ability students.

TABLE 1

Sample Sizes by Gender, Ethnicity, and Ability for Spiral 1, Study 1

	Male			*Female*		
Ethnic Group	*<410*	*410–600*	*>600*	*<410*	*410–600*	*>600*
African Am.	76	168	16	137	326	36
Asian Am.	66	299	98	100	320	106
Mexican Am.	29	98	15	70	131	17
Puerto Rican	6	26	2	7	46	6
Other Latino	19	74	12	29	129	18
White	156	1,791	665	256	2,489	746
Total	352	2,456	808	599	3,441	929

TABLE 2

Sample Sizes by Gender, Ethnicity, and Ability for Spiral 3, Study 1

	Male			*Female*		
Ethnic Group	*<410*	*410–600*	*>600*	*<410*	*410–600*	*>600*
African Am.	92	143	16	142	313	28
Asian Am.	18	158	200	29	300	170
Mexican Am.	21	79	23	65	140	19
Puerto Rican	9	24	6	12	34	2
Other Latino	18	78	19	41	99	17
White	141	1,471	838	294	2,333	621
Total	299	1,953	1,102	583	3,219	857

Table 2 contains comparable information for spiral 3 (a mathematics spiral) with the ability groupings based on operational mathematics scores rather than verbal scores. The greatest contrast with Table 1 is in the Asian American group, in which the proportion of examinees in the highest score band is substantially higher for math scores than for verbal scores.

Results and Discussion

Study 1 Verbal Item-Level Analyses

The proportion correct for the 23 V1 questions that were common to spirals 1 (standard length), 5 (8 items shorter), and 9 (12 items shorter) are shown in Figure 1. Over most of the test, the proportion correct for each common item was nearly identical across spirals, though for the last three common items the proportion correct was somewhat higher in the two shorter spirals; the proportion correct on the final items was no higher in the 23-item spiral than in the 27-item spiral.

Figure 2 shows the proportion of examinees in each V1 spiral who did not respond to an item either

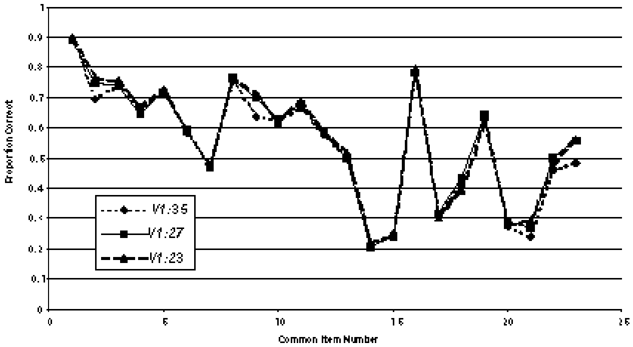

Figure 1. Proportion correct for the 23 common V1 items under standard and two less speeded conditions.

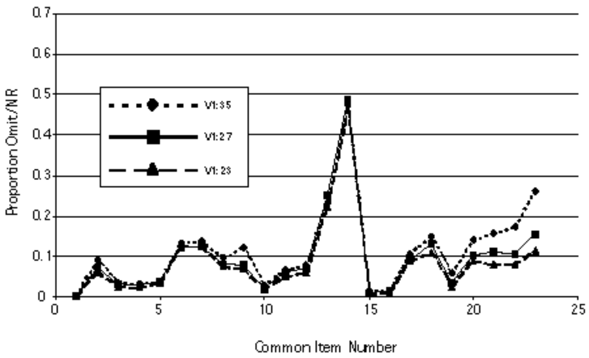

Figure 2. Proportion of examinees omitting or not reaching an item for the 23 common V1 items under standard and two less speeded conditions.

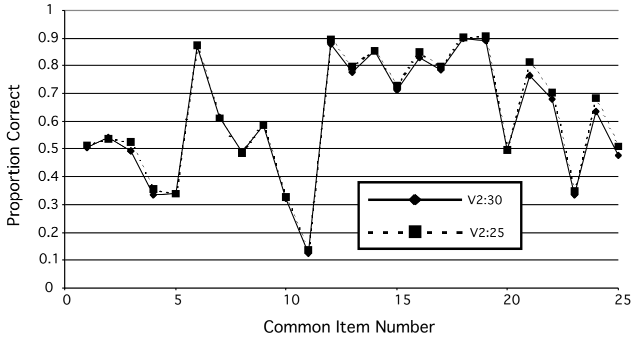

Figure 3. Proportion correct for the 25 common V2 items under standard and less speeded conditions.

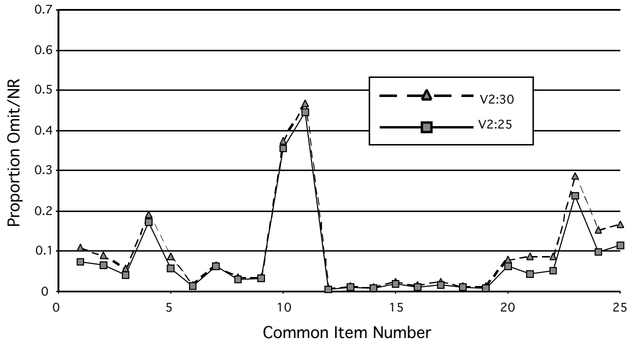

Figure 4. Proportion of examinees omitting or not reaching an item for the 25 common V2 items under standard and less speeded conditions.

TABLE 3

Percent Attempting Final Verbal Item and Percent Correct Among Those Attempting Final Item

V1			*V2*		
Length	*% Attempts*	*% Correct*	*Length*	*% Attempts*	*% Correct*
35	74	66	30	83	57
27	84	66	25	89	58
23	89	63			

because they ran out of time or chose not to answer. Because the SAT is a formula-scored test in which wrong answers carry a greater penalty than omitted answers, omitting cannot be equated with running out of time. Although items are sometimes labeled as "not reached" if the examinee does not attempt to answer any subsequent questions, there is no way of distinguishing an item that was truly not reached from an item that was intentionally omitted. The high omit rate for common item 14 appears to be primarily a function of item difficulty, not running out of time, because this item appears relatively early in the verbal test and the omit rate is fairly comparable across the three timing conditions. However, a differential omitting rate by timing condition was evident as early as common item 18 and was quite noticeable by item 20.

The proportion correct for the common items in V2 is presented in Figure 3, and the proportion omitting each item is in Figure 4. Although there is some evidence for the graphs diverging as early as common item 21, differences were small with respect to both proportion correct and omits. The standard-length V2 is five items shorter than the standard-length V1, albeit with a higher proportion of the items based on reading a passage, and V2 appears to be less speeded as indexed by the differences between speededness groups in proportion correct and proportion of omits on the last item.

Though the differences are not large, it is clear from Figures 2 and 4 that more examinees attempt to respond to the last few items when they have more time. If the additional time merely allowed lower-ability students to attempt questions that they could not answer, then the percent correct should decline as the percent attempted increased. Table 3 addresses this issue.

With a modest increase in time per item (V1 length 27 and V2 length 25), the percent correct among those attempting the last item remained just as high even though there were more attempts. However, with the more generous limits reflected in length 23, there was a drop in the success rate of those attempting the item.

Study 1 Math Item-Level Analyses

Figures 5–8 parallel Figures 1–4, except for math instead of verbal items. Differences were much more dramatic on the math items, especially for M1. Although strict time limits might be expected to impact performance on the last few items on a test, what is surprising in the M1 figures is how early in the test the groups diverged. By common item 5 there were already noticeable differences, and by item 10, six percentage points separated the standard and least speeded groups on both the proportion correct and proportion omitted graphs. Although these items occurred early in the test, it could still be the case that students in the less speeded groups moved at the same pace as students in the standard timing condition but then had time at the end of the test to revisit earlier items that they had initially skipped. Thus, "early" in the way items were presented may not necessarily be "early" in when they were actually answered.

Because the last M1 item was quite difficult, and because having an extra few minutes does not confer any additional mathematical abilities, the differences in the percent correct across groups for the last item was muted, though it was still eight percentage points. However, on the next-to-last common item (which was an easier item), the difference was 15 percentage points.

Table 4 shows the percent correct among the students who attempted to answer the last math question. Despite the large increase in the percentage of students attempting the last M1 item (from 43 percent to 70 percent), the percent correct among those attempting the item remained stable, suggesting that the test is speeded even for the high-ability students who could answer this difficult question (when they had time to consider it).

The largest gain from more time per item was found for the next-to-last common item in M1. Figures 9 and 10 show this gain for gender and ethnic groups. The gain was about the same for males and females. Although all ethnic groups showed substantial gains, the largest gain was in the sample of white students. Though gains were somewhat smaller for the M2 items, they showed the same gender and ethnic pattern. These data, then, would not support the notion that allowing more time would be likely to narrow subgroup differences.

Item-level results for Study 2 told essentially the same story. They are presented in the Appendix.

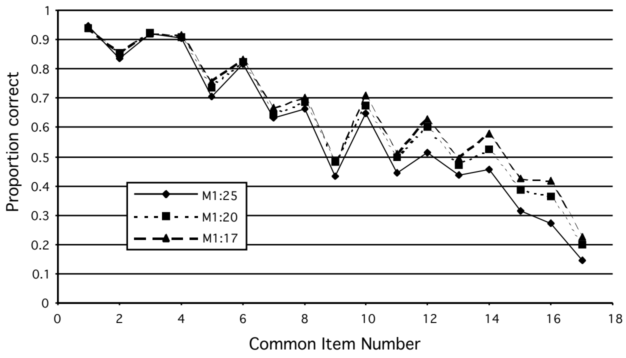

Figure 5. Proportion correct for the 17 common M1 items under standard and two less speeded conditions.

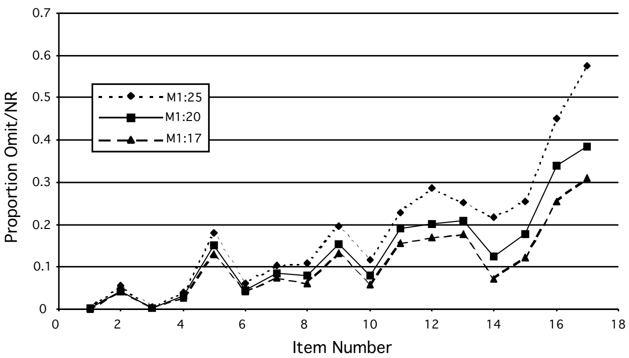

Figure 6. Proportion of examinees omitting or not reaching an item for the 17 common M1 items under standard and two less speeded conditions.

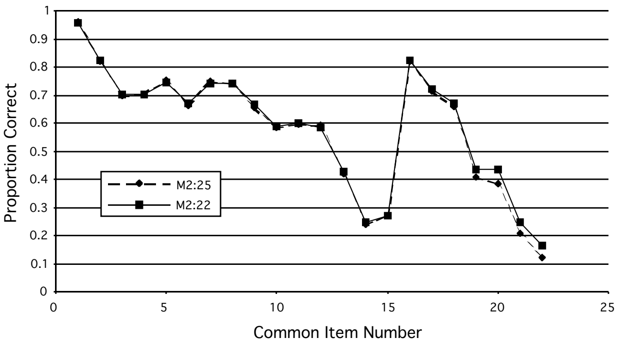

Figure 7. Proportion correct for the 22 common M2 items under standard and less speeded conditions.

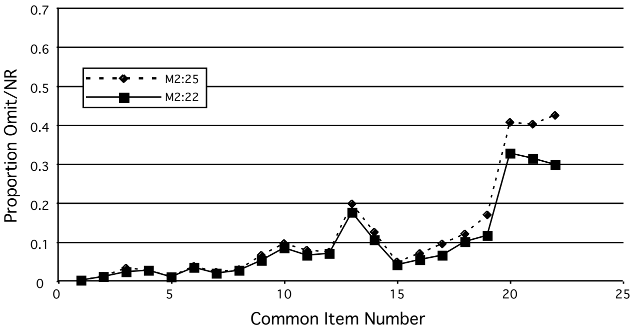

Figure 8. Proportion of examinees omitting or not reaching an item for the 22 common M2 items under standard and less speeded conditions.

TABLE 4

Percent Attempting Final Math Item and Percent Correct Among Those Attempting Final Item

M1			*M2*		
Length	*% Attempts*	*% Correct*	*Length*	*% Attempts*	*% Correct*
25	43	31	25	58	21
20	62	32	22	70	24
17	70	32			

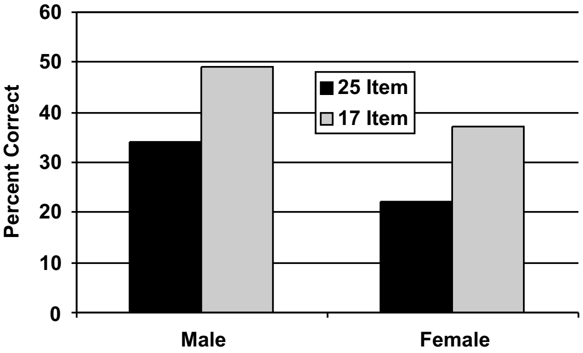

Figure 9. Percent correct, by gender, on the next-to-last M1 item.

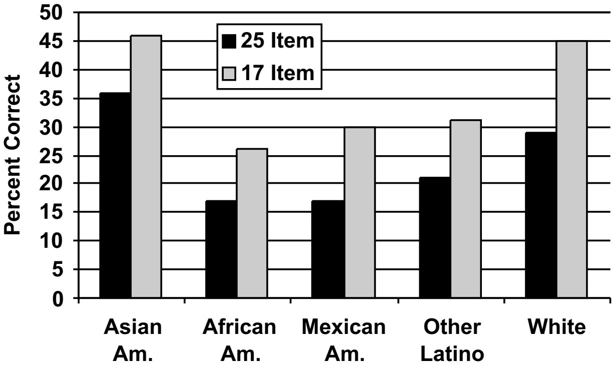

Figure 10. Percent correct, by ethnic group, for the next-to-last M1 item.

Study 1 and Study 2 Section-Level Effects—Verbal

Although the statistical analyses were run on the formula scores, we scaled the scores to the familiar 200–800 scale for presentation purposes. Specifically, the formula scores on the common items from the sections with standard timing were scaled to scores on the corresponding (verbal or mathematics) operational sections via a single group equipercentile scaling with 3 Tukey-Cureton smoothings. The relationships between formula scores and scaled scores were then applied to the sections with more generous time per item. In both studies, despite the large sample sizes, there were no statistically significant interactions (at the .05 level) of timing condition with either gender or race/ethnicity, indicating that the effects of extra time can reasonably be considered to be the same regardless of gender or racial/ethnic group. However, there were sometimes interactions with ability level. Therefore, the tables and graphs separate examinees by ability level, but not by gender or race/ethnicity. Table 5 and Figure 11 show the scaled scores for V1 in the three ability strata. In both studies, the benefits of extra time were minimal—less than eight points on the 200–800 scale in any ability level.

Results for V2 are found in Table 6 and Figure 12. Compared to the results for V1, the overall effect size for V2 was similar, with the largest difference being 10 points.

Best Language. Effects of extra time per item might be expected to be larger for students whose best language was not English. Examinees who complete the Student Descriptive Questionnaire when they register to take the SAT are asked to indicate their best language; the options are a) English, b) English and another language, and c) another language. Students who selected option “a” were categorized as English best and were compared to students in the other two categories combined. In each spiral, between 700 and 800 students were in the other two categories. Surprisingly, for both V1 and V2, statistical tests for the interaction of language and spiral were not significant (*F*s from the ANOVA less than one in Study 1 and less than 1.5 in Study 2), indicating that there was no support for the hypothesis that gains would be greater for students whose best language was not English.

Table 5

Ns, Means, and SDs for the 23 Common V1 Items on SAT Scale

		Section Length									*Mean Score*
	Ability	*35-Items*			*27-Items*			*23-Items*			*Difference*
Study	*Level*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*Shortest-Longest*
1	<410	1,156	388	67	1,160	389	68	1,046	387	68	-1
	410–600	7,102	510	72	6,791	516	74	6,480	514	76	4
	>600	2,110	641	63	2,071	644	60	1,977	642	60	1
2	<410	1,625	390	71	1,574	395	74	1,505	396	74	6
	410–600	7,272	507	76	7,169	510	77	6,846	514	77	7
	>600	2,276	642	69	2,174	646	69	2,108	649	68	7

Table 6

Ns, Means, and SDs for the 25 Common V2 Items on SAT Scale

		Section Length						*Mean Score*
	Ability	*30-Items*			*25-Items*			*Difference*
Study	*Level*	*N*	*M*	*SD*	*N*	*M*	*SD*	*Shortest-Longest*
	<410	1,019	381	64	1,258	391	63	10
1	410–600	5,987	511	72	7,202	516	70	5
	>600	1,934	644	62	2,315	648	64	4
	<410	1,386	383	64	1,273	388	66	5
2	410–600	6,552	510	73	6,215	517	73	7
	>600	2,123	643	64	1,914	646	64	3

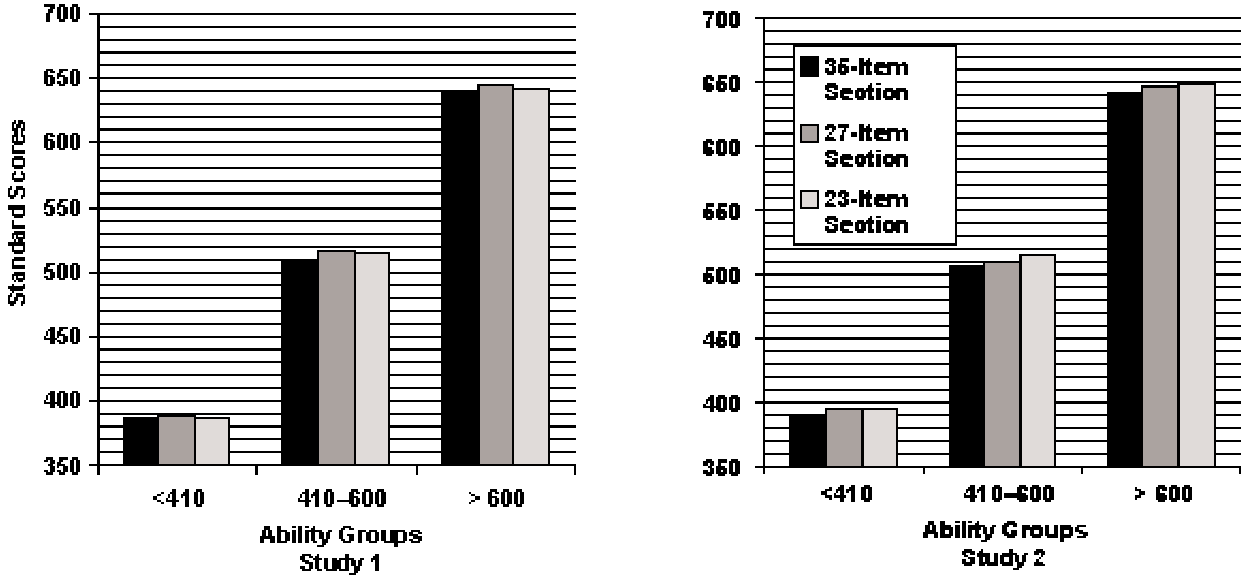

Figure 11. Mean scores on 23 V1 items with standard timing (embedded in a 35-item section), and with two less speeded conditions (embedded in a 27-item section and as a complete 23-item section).

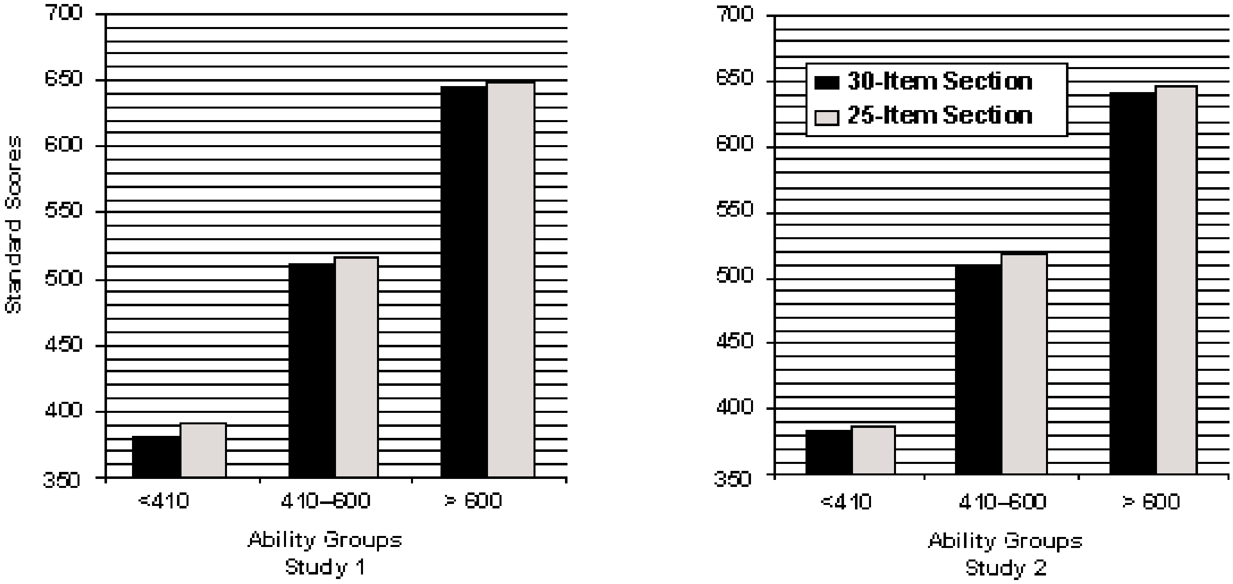

Figure 12. Mean scores on 25 V2 items with standard timing (embedded in a 30-item section), and with a less speeded condition (a complete 25-item section).

Study 1 and Study 2 Section-Level Effects—Math

Scaled scores for M1 are presented in Table 7 and Figure 13. Both Studies 1 and 2 show that for examinees whose operational scores were 400 or lower extra time was of little or no benefit on average. If a student lacks the skills to approach a problem, providing extra time will not help. Extra time is beneficial only if a student has a solution strategy, but does not have time to fully implement that strategy. In the higher-ability groups extra time was clearly beneficial, but Study 1 showed slightly larger gains than Study 2.

Given the trend of greater gains as ability increases, gains might be expected to be quite large at the highest ability levels. However, if a student is already getting nearly all of the items correct under standard timing, there is only a limited opportunity to get a higher score with more time. In Study 1 there were about 470 examinees in each spiral with operational math scores in the 700–750 score range. For these high-ability students, there was a 15 point benefit of extra time (comparing the scores on the common items from the 25-item and 17-item sections). This was less than the 26 points in the broader 600 or above range. In Study 2 the comparable difference was 12 points.

Results for M2 are presented in Table 8 and Figure 14. As with M1, there were little or no gains for the lowest ability group. Although only 3 items were eliminated to create the shorter version of M2, these items were all of the presumably more time- consuming type in which examinees have to grid in a numerical answer rather than selecting among answer choices. These results suggest that examinees, especially at the higher ability levels, could benefit from extra time on M1, but that a modest time extension would be of limited value on M2.

Best Language. For both M1 and M2, statistical tests of the language by spiral interaction were not significant (ANOVA *F*s less than one in Study 1 and also for M2 in Study 2). In Study 2 the interaction for M1 was statistically significant (*F* = 3.67 [2, 26,689], p = .03), but of no practical importance. Students whose best language was English gained 13 points from the 25-question section to the 17-question section while students for whom English was not their best language gained 7 points. Allowing more time should have little or no differential impact on mathematics scores for students whose best language is not English.

Validity of Less Speeded Tests

To the extent that speed is not part of the construct that the SAT is intended to assess, validity could increase as speededness decreases. On the other hand, there may also be cases in which more time permits students to use strategies (such as working backwards from the answer choices) that could result in a poorer assessment of their mathematical reasoning skills. One aspect of validity that we could assess was the relationship of the more and less speeded tests with external criteria. Predictive validity information relating scores to college grades was not available. It was possible, however, to evaluate the relationship of test scores with grades in high school mathematics courses using the self-reported grades that students provide on the Student Descriptive Questionnaire that they fill out when they register to take the SAT. Specifically, students are asked to enter the average grade for all courses already taken in mathematics. In another question, they are asked to indicate the total number of years "you have taken or plan to take in the specific courses listed." We divided the courses into three levels; the first level was for students who had not taken (and did not plan to take) any trigonometry or precalculus, the second level was for students who had taken (or planned to take) trigonometry or precalculus but not calculus, and the third level was for students who had taken or planned to take calculus. Correlations of test scores and math grades are summarized in Table 9. Standard errors of these correlations were about 0.02, so less speeded tests were not

TABLE 7

Ns, Means, and SDs for the 17 Common M1 Items on SAT Scale

		Section Length									*Mean Score*
	Ability	*25-Items*			*20-Items*			*17-Items*			*Difference*
Study	*Level*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*Shortest-Longest*
	<410	1,087	388	69	985	388	71	930	384	72	-4
1	410–600	6,297	512	74	5,890	523	80	5,657	535	82	23
	>600	2,390	654	66	2,372	671	65	2,247	680	62	26
	<410	1,251	381	71	1,281	384	72	1,279	386	75	5
2	410–600	6,799	514	75	6,533	522	80	6,189	529	80	15
	>600	2,995	643	66	2,819	657	66	2,811	660	67	17

Table 8

Ns, Means, and SDs for the 22 Common M2 Items on SAT Scale

		Section Length						*Mean Score*
	Ability	*25-Items*			*22-Items*			*Difference*
Study	*Level*	*N*	*M*	*SD*	*N*	*M*	*SD*	*Shortest-Longest*
1	<410	1,203	376	66	1,130	377	65	1
	410–600	6,798	513	72	6,505	516	75	3
	>600	2,638	647	64	2,522	653	65	6
2	<410	1,162	378	60	1,074	376	60	-2
	410–600	5,953	512	72	5,529	517	74	5
	>600	2,552	647	66	2,383	657	68	10

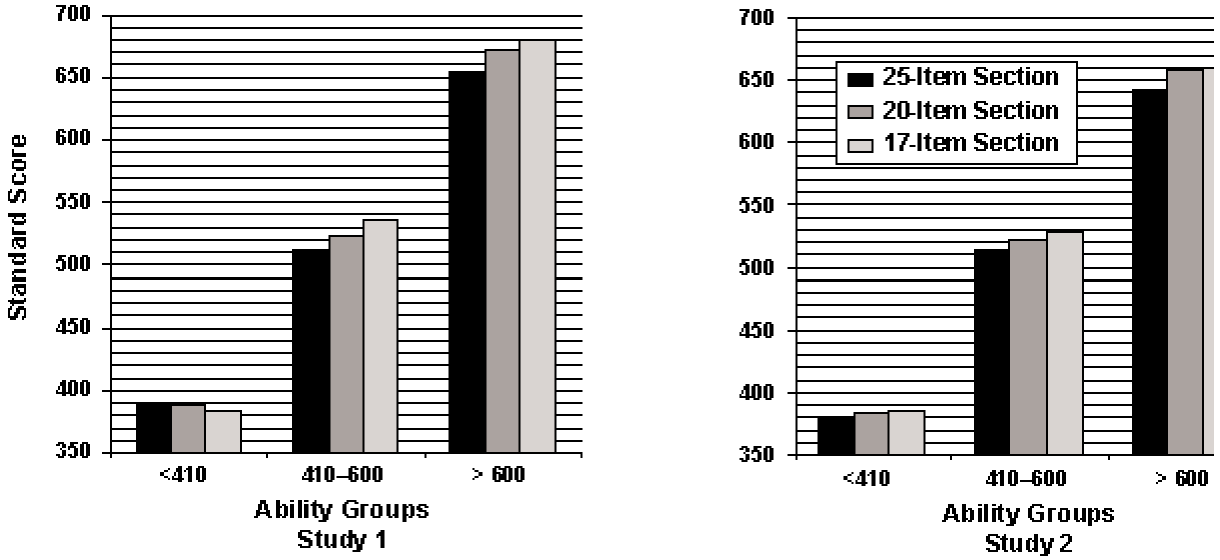

Figure 13. Mean scores on 17 M1 items with standard timing (embedded in a 25-item section), and with two less speeded conditions (embedded in a 20-item section and as a complete 17-item section).

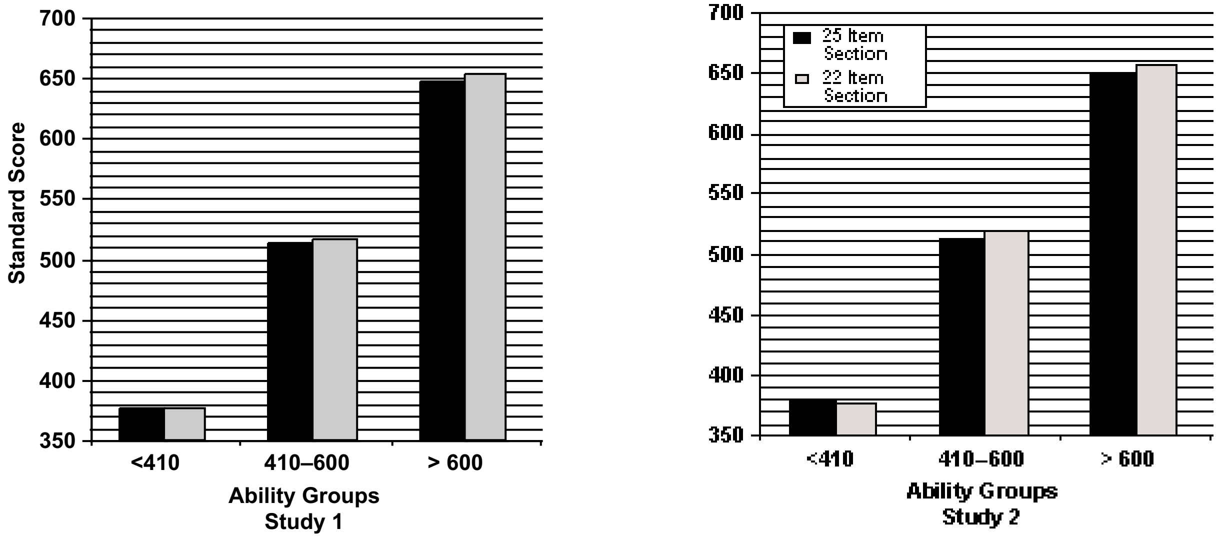

Figure 14. Mean scores on 22 M2 items with standard timing (embedded in a 25-item section), and with a less speeded condition (a complete 22-item section).

Table 9

Correlation of M1 with Math Grade

# scored/ section length	Highest Math Course					
	Below Precalc.		Precalc.		Calculus	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
25/25	.42	.43	.36	.40	.34	.34
17/25	.42	.41	.35	.39	.33	.32
17/20	.41	.43	.37	.35	.36	.34
17/17	.38	.40	.37	.40	.31	.36
20/25	.45	.41	.38	.39	.34	.33
20/20	.43	.44	.37	.36	.37	.36

Note: Minimum sample sizes in Study 1 for the Below Precalc/trig, Precalc/trig, and Calculus correlations respectively were 1,923; 3,377; and 2,099; sample sizes in Study 2 were comparable.

Table 10

Correlation of V1 with Full SAT I Verbal

# scored/ section length	Study 1	Study 2
35/35	.87	.85
23/35	.83	.82
23/27	.84	.81
23/23	.84	.81
27/35	.85	.82
27/27	.86	.81

Table 11

Correlation of V2 with Full SAT I Verbal

# scored/ section length	Study 1	Study 2
30/30	.87	.87
25/30	.85	.85
25/25	.85	.85

Table 12

Correlation of M1 with Full SAT I Math

# scored/ section length	Highest Math Course					
	Below Precalc.		Precalc.		Calculus	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
25/25	.81	.82	.83	.83	.85	.84
17/25	.76	.78	.79	.79	.81	.79
17/20	.79	.79	.81	.78	.81	.77
17/17	.80	.80	.81	.79	.78	.76
20/25	.79	.81	.81	.82	.83	.81
20/20	.82	.82	.82	.81	.83	.79

Table 13

Correlation of M2 with Full SAT I Math

# scored/ section length	Highest Math Course					
	Below Precalc.		Precalc.		Calculus	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
25/25	.83	.84	.82	.84	.83	.83
22/25	.81	.82	.80	.83	.81	.82
22/22	.81	.84	.82	.83	.81	.81

systematically more or less valid than the more speeded tests. Similar nonsignificant differences were found for M2.

Verbal test scores were correlated with high school English grades. No significant differences across groups were found. Correlations ranged from .33 to .34 for V1 and both correlations for V2 were .36.

Reliability of Less Speeded Tests

Internal consistency reliability measures, such as KR-20 or coefficient alpha, are known to artificially inflate estimates for speeded tests. Thus, comparing more and less speeded tests on such indices is problematic. Instead, we estimated pseudo-reliability coefficient by correlating scores from the various spirals with corresponding scores from the operational portion of the test. (We call this pseudo-reliability because tests of unequal length rather than truly parallel forms are being correlated.) To the extent that these operational scores were influenced by a speed component, there could be a slight bias in favor of finding higher correlations for the spirals that shared this speed component. However, despite this potential bias, correlations with operational scores were as high in the less speeded spirals as in the spirals that had the same time constraints as the operational sections. Tables 10 and 11 show these correlations for the verbal spirals, demonstrating that the less speeded spirals are as reliable as the more speeded spirals when the number of items is held constant.

Tables 12 and 13 show comparable correlations for the mathematics spirals, separately for students in the three levels of high school courses identified above. Again, there are no consistent differences among the different degrees of speededness.

Conclusions

These two studies suggest that SAT I Verbal is only slightly speeded. On both sections (V1 and V2) and in all three ability groups in both studies, the equivalent of time-and-a-half raised scores by no more than 10 points on the 200–800 scale. SAT I Math appears to be more speeded but not highly speeded; the equivalent of time-and-a-half raised scores about 20 points, though the size of the increase was somewhat larger (17 to 26 points) for higher ability students; extra time was of absolutely no benefit for students in the 400 and below score range. Consistent with previous research (e.g., Wild, Durso, and Rubin, 1982; Sackett et al., 2001), test speededness does not appear to contribute to ethnic/racial and gender differences, so creating a less speeded SAT I would have little or no impact on group differences. A shorter SAT I Mathematics test (allowing more time per item) should not have a noticeable impact on validity, at least to the extent that this can be estimated from correlations with concurrent math grades.

A possible limitation of the current studies is that students had no advance notice that they would be taking a section with more generous time limits, so they never had the opportunity to practice at the more relaxed pace permitted by the shorter sections. Students who worked at their standard pace on the shorter sections would have had more time at the end to review and revise their previous answers, but this may not be equivalent to working at a slower pace throughout the test.

A less speeded mathematics test could provide a number of potential benefits. First, it would be a better representation of the mathematics construct that the test is designed to assess in which speed of performance is expected to play a minor role in determining scores (Donlon, 1984). Second, a less speeded test is desirable now that scores of disabled students who are granted extra time will no longer be flagged. The flag had been a signal that scores from nonstandard administrations may not be comparable to scores from standard administrations. A flag is not needed and not used for accommodations, such as large print for visually impaired students, that would not impact the scores of nondisabled students even though they have a large impact for students that need them. If time limits were sufficient so that extended time provided a trivial impact on scores for nondisabled students, there would not be a reason to flag scores of extended-time administrations. The current studies did not include disabled populations and so must be silent on the question of how much extra time increases scores for the disabled. But the argument to not flag is more dependent on showing that extra time is of minimal benefit for the nondisabled population. With a less speeded test, the pressure for students to get a sometimes questionable diagnosis in order to qualify for extra time would be substantially reduced, as would the pressure on the College Board to determine whether those diagnoses were legitimate. Removal of the necessity for a flag also would benefit the truly disabled, who would then no longer have to involuntarily disclose their status. Third, more generous time limits would have a positive impact on test-preparation activities that could focus on problem-solving strategies rather than strategies aimed largely at beating the clock.

References

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. *American Educational Research Journal, 27*, 65–87.

Camara, W. J., Copeland, T., & Rothschild, B. (1998). *Effects of extended time on the SAT I: Reasoning test score growth for students with disabilities* (College Board Report No. 98-7). New York: College Entrance Examination Board.

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

Evans, F. R. (1980). *A study of the relationships among speed and power aptitude test score, and ethnic identity* (ETS RR 80-22). Princeton, NJ: Educational Testing Service.

Linn, M. C. (1992). Gender differences in educational achievement. In *Sex equity educational opportunity, achievement, and testing: Proceedings of the 1991 ETS Invitational Conference* (pp. 11–50). Princeton, NJ: Educational Testing Service.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*, 261–270.

Sackett, P. R., Schmitt, N., Ellingston, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56*, 302–318.

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

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effects of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement, 19*, 19–28.

Appendix:
Item-Level Results
for Study 2

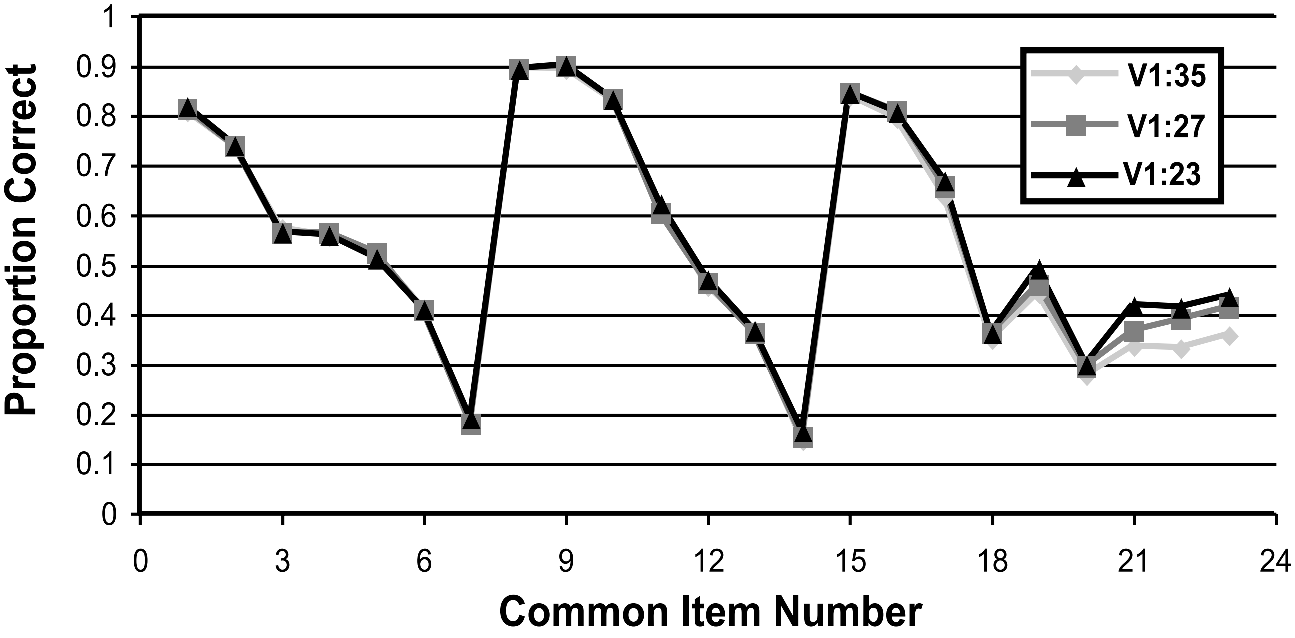

Figure A1. Proportion correct for the 23 common V1 items under standard and two less speeded conditions.

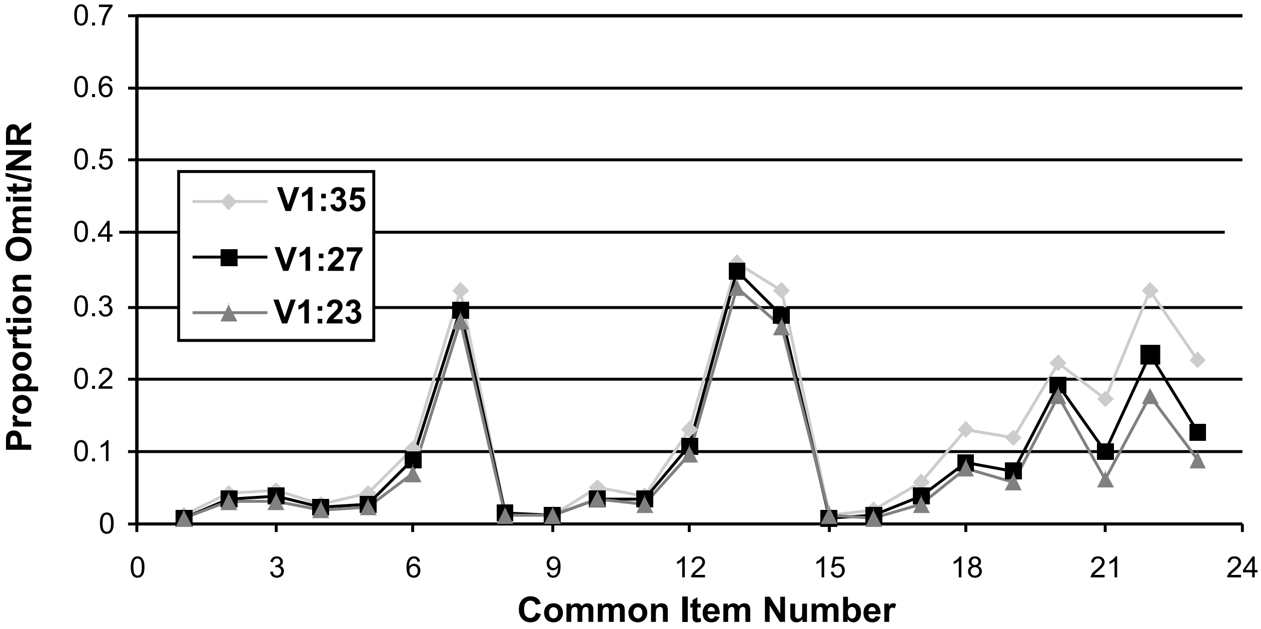

Figure A2. Proportion of examinees omitting or not reaching an item for the 23 common V1 items under standard and two less speeded conditions.

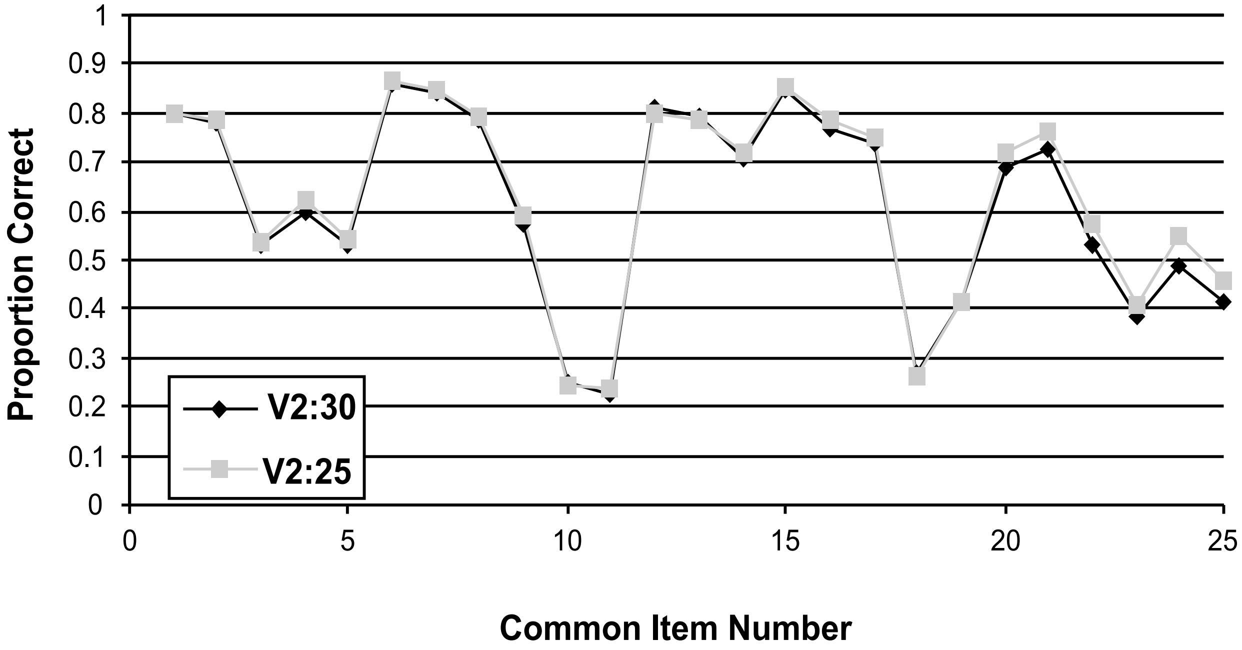

Figure A3. Proportion correct for the 25 common V2 items under standard and less speeded conditions.

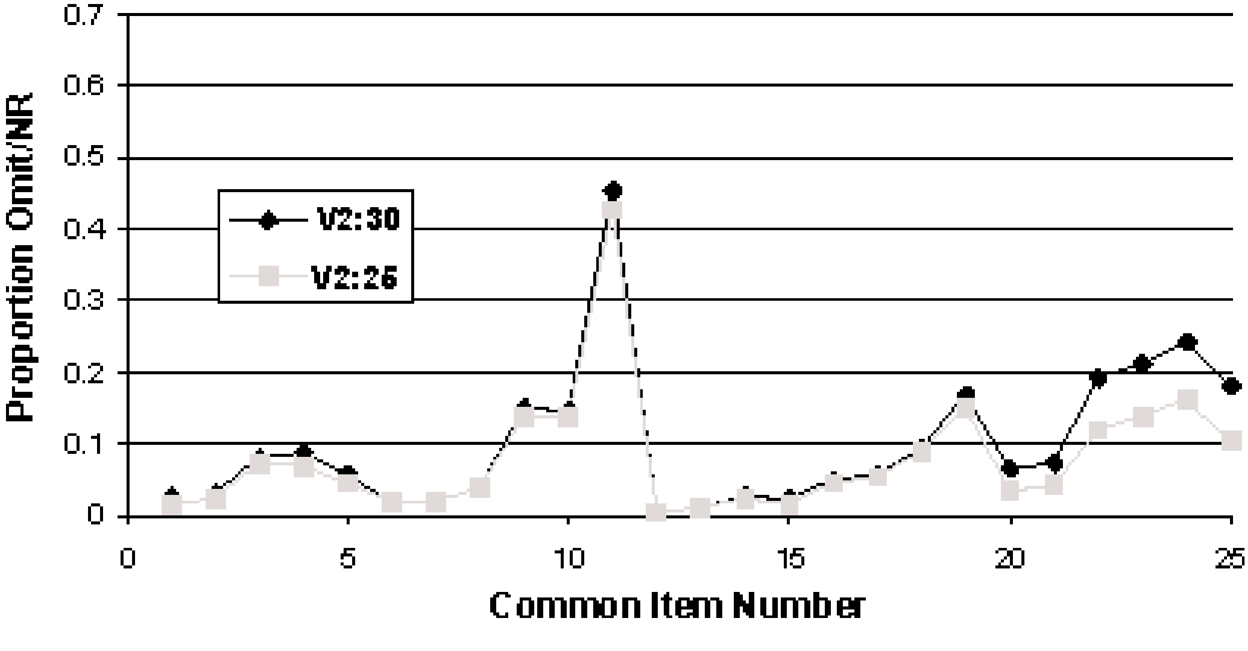

Figure A4. Proportion of examinees omitting or not reaching an item for the 25 common V2 items under standard and less speeded conditions.

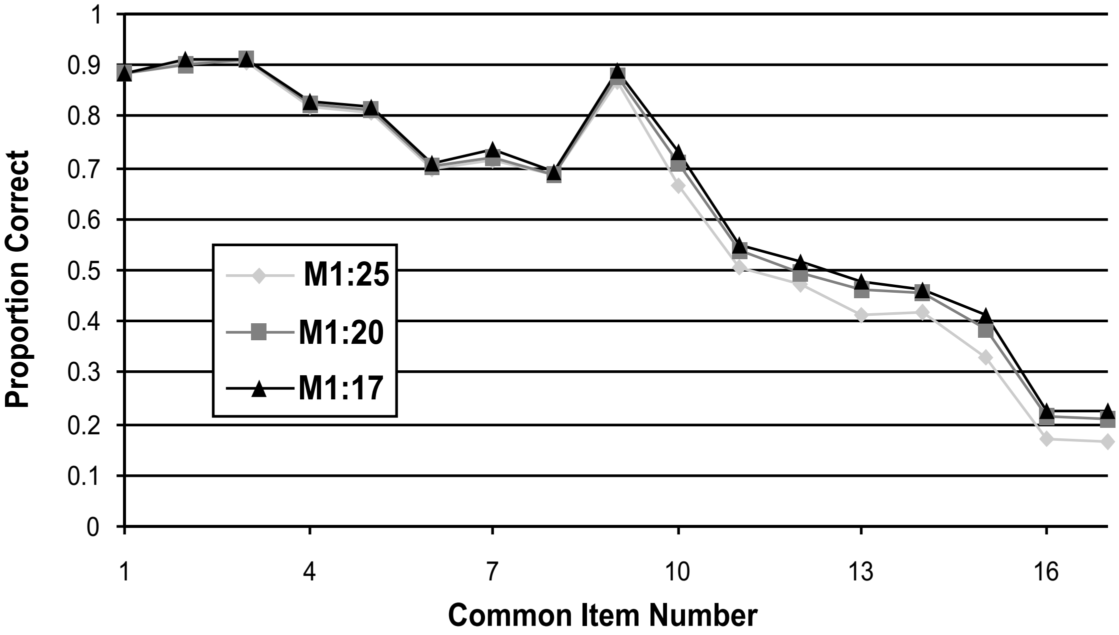

Figure A5. Proportion correct for the 17 common M1 items under standard and two less speeded conditions.

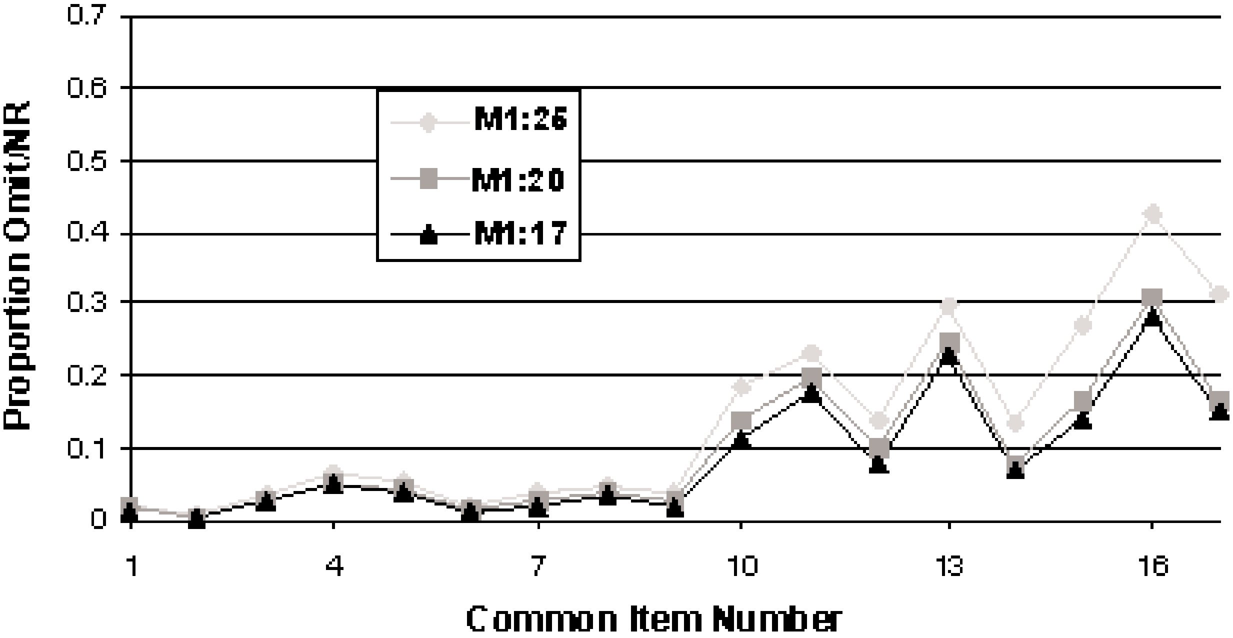

Figure A6. Proportion of examinees omitting or not reaching an item for the 17 common M1 items under standard and two less speeded conditions.

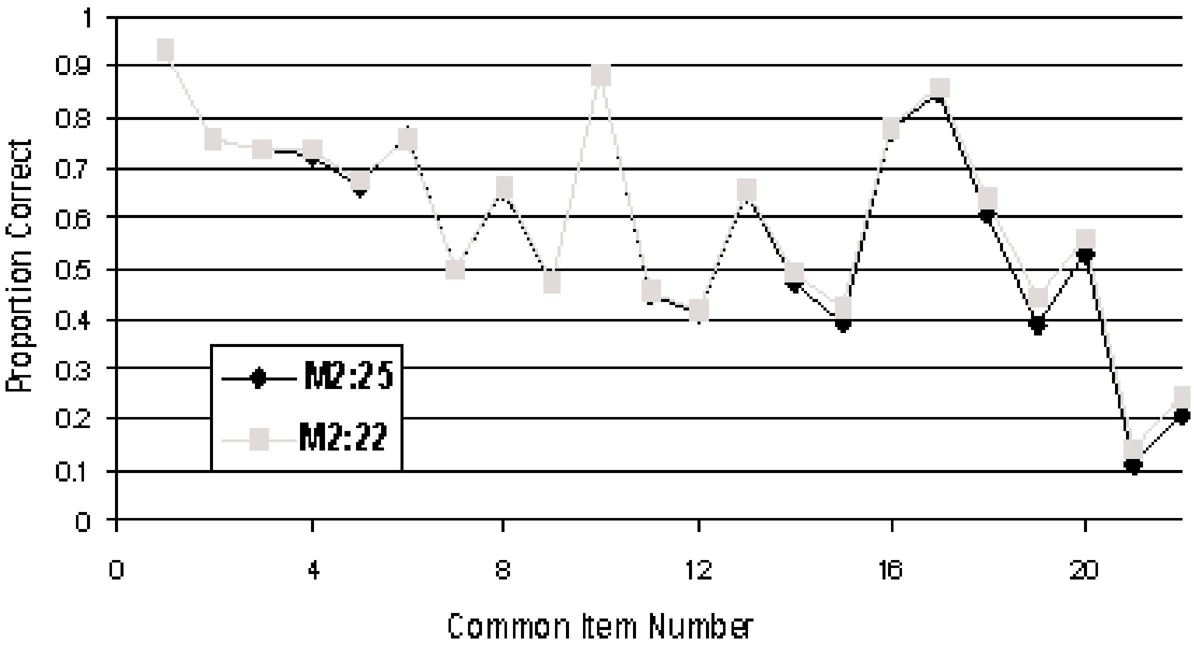

Figure A7. Proportion correct for the 22 common M2 items under standard and less speeded conditions.

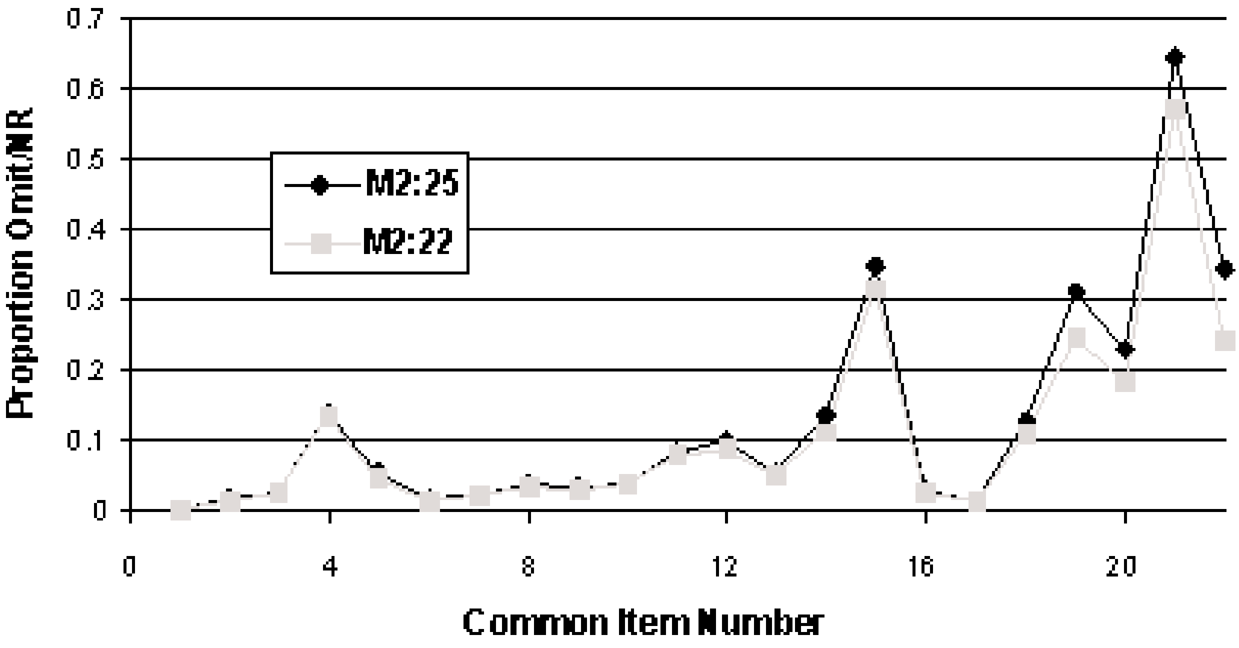

Figure A8. Proportion of examinees omitting or not reaching an item for the 22 common M2 items under standard and less speeded conditions.

www.collegeboard.com