R&D Connections
November 2004

Testing and Time Limits

Brent Bridgeman, Amanda McBride, & William Monaghan

Testing and time limits. It's an almost inevitable union—and for good reason, many would argue. Imposing time limits on tests can serve a range of important functions. Time limits are essential, for example, if speed of performance is an integral component of what is being measured, as would be the case when testing such skills as how quickly someone can type.

Limiting testing time also helps contain expenses associated with test administrations, such as paying hourly fees for proctors in a paper-based administration or for seat time at computer testing centers.

But limiting testing time too drastically can threaten a test's validity, or the ability of the test to accurately reflect what the test was designed to measure. This is particularly true if the test is not intended to measure how quickly the test taker can answer questions or if the testing time is so limited that a large number of examinees taking the test cannot complete it; that is, if the test is "speeded."

Speededness in testing refers to the effect that time limits have on test takers' scores. When a test's time limits are constrained to the point that most test takers do not have enough time to consider and answer each question, the test is said to be "speeded." A test is speeded to the extent that those taking it score lower than they would have if they had been given an unlimited amount of time to complete it.

For tests such as the GRE® and College Board's SAT®, which are intended to measure skills related to academic ability rather than the rate at which examinees can work, the speed at which test takers answer the questions should play a minor role, at most, in determining test scores (Briel, O'Neill, & Scheuneman, 1993; Donlon, 1984). Consequently, time limits for such tests should give most test takers enough time to finish the test, and a modest time extension should have a relatively small effect on overall test scores (Bridgeman, Cline, & Hessinger, 2003).

While it's possible that time limits can affect the scores of all test takers, some have suggested that such limits may differentially affect female and minority test takers. Some claim that the "fast-paced, or speeded nature" of the SAT puts female test takers at a disadvantage on certain test sections because they approach problem-solving differently than their male counterparts—female test takers, they say, are more likely to work problems out completely, to consider more than one possible answer, and to check their work (Becker, 1990; Linn, 1992).

Others have noted what seems to be a common belief among test takers and their families (and even among some school counselors) that giving examinees more time to complete a test could substantially improve their scores. This has raised concerns over the possibility that nondisabled students may attempt to obtain extended-time accommodations (which ETS provides to examinees with documented disabilities that require additional testing time, such as learning disabilities, Attention-Deficit/Hyperactivity Disorder, or sight problems), and thus gain a perceived advantage on standardized tests (Bridgeman, Trapani, & Curley, 2003; Mandinach, Cahalan, & Camara, 2002). But if evidence suggests that extra time does not improve test taker performance, students would have little or no motivation to manipulate the system to receive extra test-taking time that they're not entitled to. And there would be less

reason to flag[1] the scores of students who were granted extended time, a practice that has engendered fierce debate since its implementation decades ago.

Effect of Extra Time on SAT Test Scores

With all this in mind, the obvious questions seem to be, what happens when test takers are given more time to complete a standardized test? Do test takers' scores improve when they are given more time? And if so, by how much?

To begin to answer these questions, Bridgeman, Trapani, and Curley (2003) placed SAT Reasoning Test™ sections with a fewer number of questions into the standard 30-minute variable section of two national test administrations. This section does not count toward the final scores of test takers, but is used to try out new questions and to ensure that scores on new editions of the test are comparable to those on earlier editions. The researchers created the reduced number sections by deleting questions from a verbal section that contained 35 questions, to produce two sets of forms, one with 27 questions and another with 23. The scores on the 23 questions could then be compared to the scores on the same 23 questions in the sections containing the 27 or 35 questions. This was done for both the math and the verbal sections of the test.

As can be seen in Figures 1, 2, 3, and 4, the researchers found that allowing more time per question (the equivalent of time-and-a-half) had minimal impact on verbal scores, producing gains of less than 10 points on the 200-800 SAT scale. In fact, in the first study, scores for the lower ability group (those who scored below 400) actually decreased with extra time. These results suggest that the SAT verbal section is only slightly speeded. The math section appears to be more speeded than the verbal section, but not highly speeded: The equivalent of time-and-a-half raised scores about 20 points, although the increase was somewhat greater (17-26 points) for higher ability students (ability level > 600).

For both sections, increasing the time tended to benefit high-scoring students more than lower-scoring students, with extra time creating no increase in scores for students with SAT scores of 400 or lower (ability level < 410).

Moreover, racial/ethnic and gender differences neither increased nor decreased with extra time, so it appears that creating a less speeded SAT would have little or no impact on group differences. And while allowing more time for the math sections would give mathematically able students a chance to answer more questions correctly, it would not affect racial/ethnic or gender differences. Additionally, scores on the forms allowing more time per question were as correlated with high school English and math grades as were scores on the forms allowing less time per question.

A goal of the revised SAT (New SAT), which will make its debut in 2005, was to make the test even less speeded than the current version. Field trials of the New SAT suggest that this has been achieved. In this trial, identical test sections were administered with either a 25-minute or 40-minute time limit. The extra time had virtually no effect on the reading section scores and only increased scores by about 10 points (on a 200 – 800 scale) on the math section (Bridgeman, 2004).

[1] "Flagging" refers to the practice by which administrators of standardized tests place asterisks or other similar notations on the score reports of people with disabilities who take exams under certain nonstandard conditions. These conditions usually involve an accommodation on or a modification to the test and may include providing people to read the test instructions and questions aloud, large-print and Braille forms of the test, individualized administration, or extended time. Accommodations are intended to eliminate irrelevant sources of difficulty that are related to the disability but not to the construct being assessed. It's worth noting that the number of students requesting extra time has grown by about 26 percent over the past five years (Camara, Copeland, & Rothschild, 1998). It's also important to note that, as of Oct. 1, 2001, ETS no longer flags scores of tests that were administered under an accommodation of extended time.

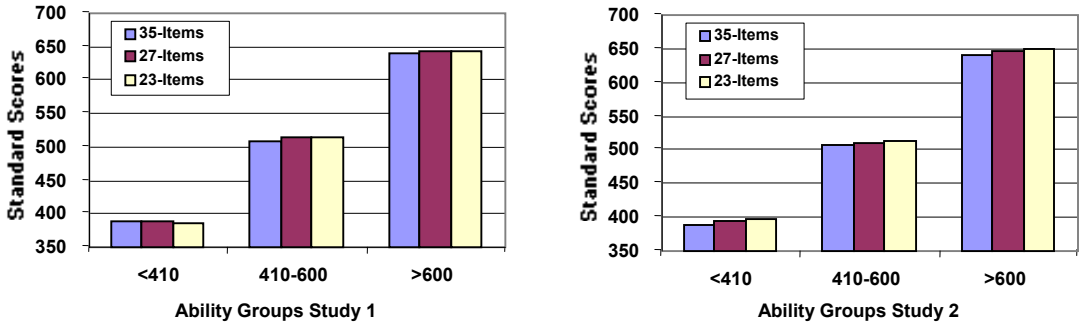

Figure 1. Mean scores on 23 V1 items with standard timing (embedded in a 35-item section), and with two less speeded conditions (embedded in a 27-item section and as a complete 23-item section).

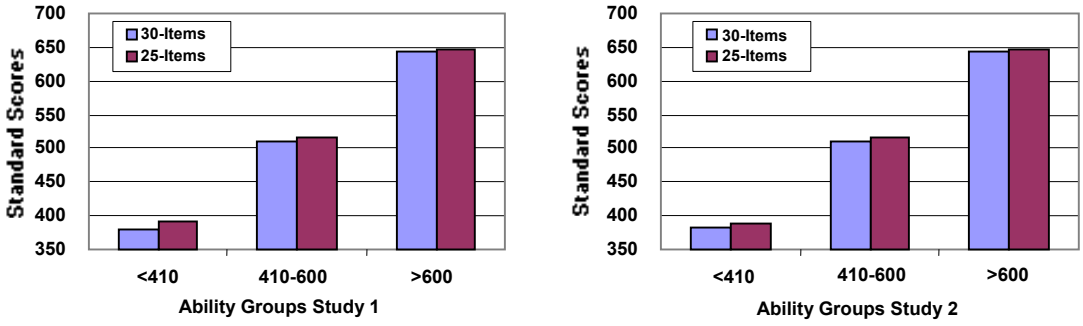

Figure 2. Mean scores on 25 M1 items with standard timing (embedded in a 30-item section), and with a less speeded conditions (a complete 25-item section).

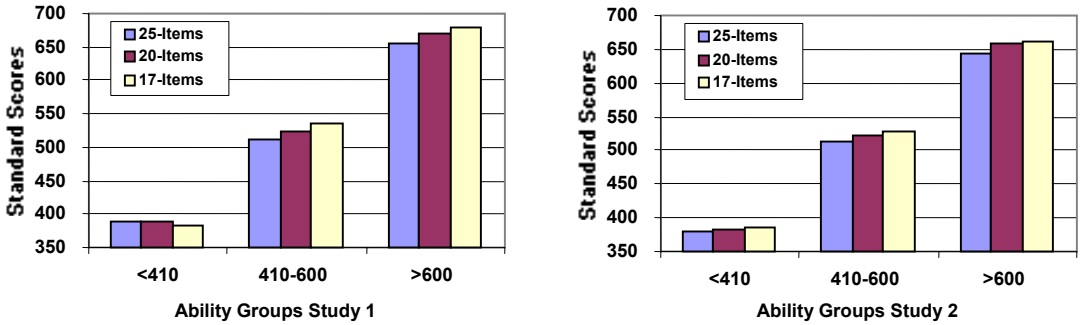

Figure 3. Mean scores on 17 M1 items with standard timing (embedded in a 25-item section), and with two less speeded conditions (embedded in a 20-item section and as a complete 17-item section).

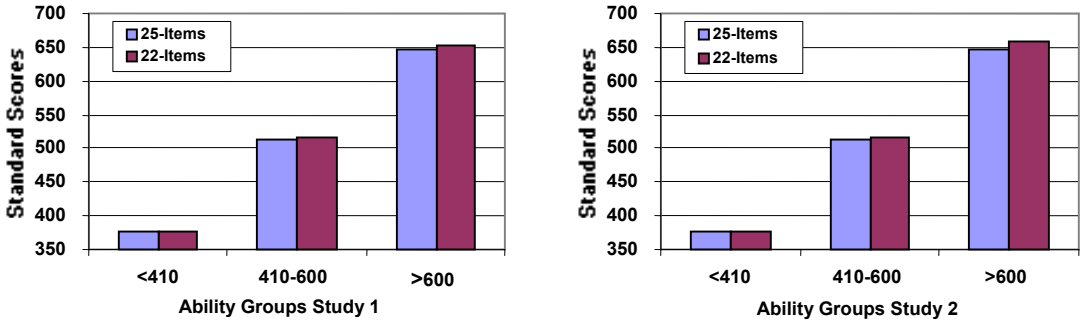

Figure 4. Mean scores on 22 M2 items with standard timing (embedded in a 25-item section), and with a less speeded condition (a complete 22-item section). *Source:* Bridgeman, Trapani, & Curley, 2003.

Effect of Extra Time on Quantitative and Verbal GRE Scores

As with the SAT, time limits for the GRE General Test are intended to be set so that most test takers can complete the test. A modest time extension, then, should have a relatively small effect on test scores. The results from the SAT study, however, cannot be applied to the current computer-adaptive GRE General Test because of the content and timing differences of the two tests, and because of the differences between computer-adaptive testing (CAT) and paper-based administration. In a CAT, unlike paper-based tests, different examinees receive different sets of questions.[2]

Unlike many CATs, the GRE CAT has a fixed number of questions and strict time limits for each section, although it is not intended to be a speeded test. To investigate speededness and the GRE CAT, Bridgeman, Cline, and Hessinger (2003) performed a study in which a research section was added to the end of regular administrations of the CAT GRE. Volunteers took either a verbal or a quantitative GRE section with either standard timing or one-and-a-half times the standard time limit. To encourage motivated performance, participants were eligible for a cash payment if they did as well on the experimental section as they did on the operational sections.

As Tables 1 and 2 show, results from this study indicate that extra time had a minimal effect on overall scores, adding only about 7 points to verbal scores and 7 points to quantitative scores on the 200-800 score scale. And, as was the case in the SAT study, scores under the different conditions were comparable across gender and ethnic groups, although quantitative scores were slightly higher for lower ability examinees who had more time. Note, however, that there are some important differences between the SAT and GRE. The SAT subtracts a fraction of a point for every question that is answered incorrectly, so that it is better to leave a question unanswered than to give an incorrect answer. The GRE, on the other hand, has a penalty for leaving questions unanswered at the end. Questions on the SAT are arranged for the most part to become successively more difficult. Lower ability test takers are more likely to guess and give incorrect answers to the latter set of questions, resulting in a negative effect on their scores. However, this is not true for sections with reading passages, which make up the majority of the verbal test. Order of those items is dependent upon where the topics the individual items refer to appear in the passage. On the GRE CAT, lower ability test takers would receive questions at or close to their ability level toward the end of the test, lessening their need to guess.

Impact of Time Limits on Computer-Adaptive Tests

As mentioned earlier, the GRE CAT is not intended to be a speeded test, but has a fixed number of questions and section time limits. So what happens when time limits are imposed on tests that give different questions to different examinees, particularly if questions that are supposed to be equally difficult tend to have substantial differences in the time it takes to answer them?

Bridgeman and Cline (2000) found that some of the questions in the GRE's analytical and

[2] In computer-adaptive testing, the computer selects the range of questions that is appropriate to each test taker's ability level. Test takers receive a set of questions that meet test design specifications and generally are appropriate for each test taker's performance level. Questions are chosen from a large pool of possible questions categorized by content and difficulty. (The content and types of questions are similar to that found in comparable paper-based tests.) The computer-adaptive test starts with questions of moderate difficulty. As the candidate answers each question, the computer scores the question and uses that information, as well as the candidate's responses to previous questions, to determine which question is presented next. As long as the test taker responds correctly, the computer typically selects a next question of greater difficulty. In contrast, if the test taker answers a question incorrectly, the computer typically selects a next question of lesser difficulty. Subsequent questions are presented based in part on the test taker's performance on previous questions and in part on the test design. In other words, the computer is programmed to fulfill the test design as it continuously adjusts to find questions of appropriate difficulty for test takers of all performance levels.

quantitative sections could be answered much more quickly than others. The researchers also noted that while some of this variation in response time was related to the difficulty of the questions—more difficult questions tended to take longer to answer than less difficult ones—there also was substantial variation in the time required to answer questions of roughly the same difficulty level and meeting the same content specifications.

Given these findings, it seemed conceivable that examinees receiving time-consuming tests (i.e., those who get a disproportionate number of items that take a longer-than-average time to answer) could be disadvantaged and, as a result, receive lower scores compared to test takers who get a less time-consuming test. Yet, upon further investigation, Bridgeman and Cline (2000) could find no evidence of impact on total test scores.

In a related study, however, Bridgeman and Cline (2004) did find evidence that test takers on the analytical section of the GRE were indeed affected by this combination of conditions, which resulted in test takers having to guess on the final questions in order to finish the test before running out of time. Test takers at the higher ability levels tended to guess more than those at the lower ability levels because the questions administered to higher ability examinees were typically more time-consuming. Since guessing increases the chances of answering items incorrectly (which would lower a test taker's score), these findings indicate that examinees who are administered tests with a disproportionate number of time-consuming items are likely to get lower scores than those of *comparable ability* who receive tests containing items that can be answered more quickly.

It's worth noting that the GRE's analytical section has been replaced by two essay prompts that assess analytical writing skills. Although the potential problem noted above contributed to this decision, it was not the only consideration (Bridgeman & Cline, 2004).

Implications

This research indicates that individuals taking either the SAT or the *verbal* and *math* sections of the GRE CAT have sufficient time to answer the questions.

Table 1

Sample Sizes, Means, and Standard Deviations for Research GRE Quantitative Scores

Statistic	Timing condition		Difference
	Standard (45 min.)	Extended (68 min.)	
n	3,904	3,749	
M	664	671	7
SD	125	121	

Table 2

Sample Sizes, Means, and Standard Deviations for Research GRE Verbal Scores

Statistic	Timing condition		Difference
	Standard (30 min.)	Extended (45 min.)	
n	4,197	4,098	
M	454	461	7
SD	122	120	

Source: Bridgeman, Cline, & Hessinger, 2003.

These tests are not speeded to any significant degree, and giving test takers more time to complete these items does not result in significant score gains. The score gains that were achieved (less than 10 points for the verbal section and less than 30 points for the math section, on a 200-800 scale) were extremely minor and would certainly not make or break a student's educational aspirations. Moreover, score gains were not consistent across ability levels: For these assessments, high-scoring test takers 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.

Furthermore, racial/ethnic and gender differences were neither increased nor reduced with extra time, challenging arguments that the so-called "speeded" nature of the SAT disadvantages minority and female test takers.

These results should help to reduce the motivation for students who are not disabled to

manipulate the system in an attempt to obtain unwarranted extended-time accommodations. At the same time, test users should not be overly concerned that some students might be gaining an unfair advantage in this manner, since any such advantage would likely be quite small.

Studies were conflicting regarding whether or not the Analytic section of the GRE CAT was speeded. Although the most recent study (Bridgeman & Cline, 2004) make a strong argument that the test was indeed speeded, it is now a moot point since ETS no longer administers this section. However, the information obtained in this study should prove useful to developing future CATs with strict time limits.

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.

Bridgeman, B. (2004, April). *Speededness as a threat to construct validity.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA. Retrieved Oct. 19, 2004, from the ETS Web site: http://www.ets.org/research/dload/NCME_2004-Bridgeman.pdf

Bridgeman, B. & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41,* 137-148.

Bridgeman, B., & Cline, F. (2000). *Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assessment* (ETS RR-00-7). Retrieved Oct. 19, 2004, from the ETS Web site: http://ftp.ets.org/pub/res/researcher/RR-00-07-Bridgeman.pdf

Bridgeman, B., Cline, F., & Hessinger, J. (2003). *Effect of extra time on GRE® Quantitative and Verbal scores* (ETS RR-03-13). Retrieved Oct. 19, 2004, from the ETS Web site: http://ftp.ets.org/pub/res/researcher/RR-03-13-Bridgeman.pdf

Bridgeman, B., Trapani, C., & Curley, E. (2003). *Effect of fewer questions per section on SAT® I scores* (College Board Report No. 2003-2). Retrieved Oct. 19, 2004, from the College Board Web site: http://www.collegeboard.com/research/pdf/rdcbreport20032web_23502.pdf

Briel, J. B., O'Neill, K. A., & Scheuneman, J. D. (1993*). GRE technical manual*. Princeton, NJ: ETS.

Camara, W., Copeland, T., & Rothschild, B. (1998). *Effects of extended time on the SAT: Reasoning test score growth for students with learning disabilities* (College Board Report No. 98-7). Retrieved Oct. 19, 2004, from the College Board Web site: http://www.collegeboard.com/research/pdf/rr9807_3912.pdf

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

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: ETS.

Mandinach, E., Cahalan, C., & Camara, W. (2002). *The impact of flagging on the admission process: Policies, practices, and implications* (College Board Report No. 2002-2). Retrieved Oct. 19, 2004, from the College Board Web site: http://www.collegeboard.com/research/pdf/02595020604txtcvr_11433.pdf

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001

Send comments about this publication to the above address or via the Web at:

http://www.ets.org/research/contact.html

Listening. Learning. Leading.