GRE®
RESEARCH

Effect of Extra Time on GRE® Quantitative and Verbal Scores

Brent Bridgeman
Frederick Cline
James Hessinger

May 2003

GRE Board Professional Report No. 00-03P
ETS Research Report 03-13

ETS® Educational Testing Service
Princeton, NJ 08541

Effect of Extra Time on GRE® Quantitative and Verbal Scores

Brent Bridgeman, Frederick Cline, and James Hessinger

GRE Board Report No. 00-03P

May 2003

This report presents the findings of a
research project funded by and carried
out under the auspices of the
Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

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

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Educational Testing Service
Princeton, NJ 08541

Abstract

The purpose of this study was to test the assumption that the Graduate Record Examinations (GRE®) General Test is a measure of academic reasoning abilities in which speed of responding plays at most a minor role. In addition to completing the operational GRE General Test, participants each completed a research version of either the GRE verbal or quantitative test within either the standard time limit or within one-and-a-half times the standard time limit. Scores obtained from 15,948 examinees indicate that extra time added about seven points (on the 200-800 score scale) to examinees' verbal scores and seven points to their quantitative scores. Scores under the different timing conditions were generally comparable across gender and ethnic groups, but quantitative scores were slightly higher for lower-ability examinees who had more time.

Key words: GRE General Test, response time, time limits, speededness

Table of Contents

List of Tables

Table of Figures

Introduction

Time limits on tests may serve at least two important functions. First, they may be needed if speed of performance is presumed to be related to the construct of interest. This is clearly the case on tests of such constructs as clerical coding speed, in which the task is so easy that individual differences emerge only with respect to how quickly examinees respond. Even on more complex reasoning tasks, such as performance on the Wechsler Intelligence Scales, speed of performance may be an inherent part of the tested construct. The second major reason for imposing time limits on tests is administrative convenience — or even administrative necessity if testing costs are to be controlled. Whether paying proctors in a paper-and-pencil administration or paying for seat time at a computer testing center, costs escalate rapidly if examinees are allowed unlimited time to complete a test.

The extent to which a time limit is imposed for construct reasons, as opposed to administrative reasons, is not always clear. The technical manual for the Graduate Record Examinations (GRE®) General Test indicates that speededness is a potential threat to the validity of the test because it "is intended to reflect intellectual power primarily, rather than the rate at which examinees work." A footnote then adds, somewhat ambiguously, "The assumption here is that variation among examinees in the rate at which they respond to test items constitutes an irrelevant source of difficulty in test performance. However, the capacity to work rapidly or to process information efficiently may be a relevant aspect of academic ability" (Briel, O'Neill, & Scheuneman, 1993, p. 83). Whether speededness is irrelevant or a relevant indicator of academic ability, the extent to which scores are dependent on time is of interest to potential score users.

Completion rates have sometimes been used to suggest that a test is speeded — that is, that scores on the test are influenced by the time limit. Clearly, if substantial numbers of students fail to finish a test, it is speeded, but if everyone finishes, there is no assurance that the test is not speeded. This is especially true for tests that impose no penalty for guessing (such as the paper-and-pencil version of the GRE General Test) and tests that impose a penalty for leaving questions unanswered (such as the computer-adaptive GRE General Test). Rapid-guessing behavior at the end of a test provides evidence that a test is speeded (Bejar, 1985; Schnipke & Scrams, 1997; Yamamoto, 1995), but still does not provide an estimate of how much scores would improve if time limits were lengthened.

An experimental study that randomly assigns participants to different timing conditions

can provide an estimate of the effects of extra time on test performance. One such study by Wild, Durso, & Rubin (1982) of an older paper-and-pencil version of the GRE General Test showed that allowing one-and-a-half times the standard time limit resulted in average score gains on experimental sections that correspond to about 25 points on the 200-800 GRE scale. These gains were constant across gender and racial/ethnic groups. In a recent study of the SAT® I: Reasoning test, Bridgeman, Trapani, and Curley (2002) found similar effect sizes when they experimentally manipulated the number of test items within a fixed period of time. These results also did not differ by gender or ethnic group, though the observed effects appeared to be larger for higher-ability examinees and nonexistent for lower-ability examinees. However, none of these results is directly applicable to the current, computer-adaptive GRE General Test because of content and timing differences between the computerized GRE exam and the SAT I (and earlier versions of the GRE exam) and because of differences between computer-adaptive testing (CAT) and paper-and-pencil administration. The current study was intended to fill this gap.

Method

Participants

Examinees who took the computer-adaptive GRE General Test over a two-month period were invited to participate in a research project that would require them to take an additional test section. At the end of the regular test, a screen appeared that offered an incentive not only to participate, but to perform well on this research section. The instructions stated:

> It is important for our research that you try to do your best on this section. The sum of $250 will be awarded to each of 100 individuals testing from September 1 to October 31. These awards will recognize the efforts of the 100 test takers on the research section. Only test takers who meet the following criteria will be eligible for the award. Awards will be given to those 100 test takers who score the highest on questions in the research section relative to how well they did on the preceding scored sections. In this way, test takers at all ability levels will be eligible for the award. Award recipients will be notified by mail.

A total of 29,962 examinees volunteered to participate and at least started to answer questions on the research section. However, about half spent so little time on this extra section that we assumed they did not make a serious effort. Consequently, we decided to screen out examinees who did not spend at least 30 minutes on the quantitative section (which has a standard time limit of 45 minutes) or at least 20 minutes on the verbal section (which has a standard time limit of 30 minutes). Although in making this decision we may have screened out some test takers who were simply exceptionally fast, this seemed preferable to including large numbers of unmotivated examinees in the sample.

As a check of the reasonableness of this screening, we compared the operational scores of examinees who were screened out with the scores of examinees who passed the screening. Of the 14,633 examinees who started the quantitative research section, 7,653 passed the screening. The mean operational score of those who were screened-out was 617, while mean score for the same test takers on the research section was 450. The correlation of these research and operational scores was just 0.34. The picture was quite different for the sample that passed the screening. Their mean operational and research scores were 657 and 667, respectively, and the correlation between these mean scores was .90. Thus, for examinees who passed the screening, performance on the research and operational sections was virtually indistinguishable.

Screening results for examinees who completed the verbal research section were comparable to the results found for those who took the quantitative section. Although the original verbal sample totaled 15,329, the screened-in sample numbered 8,295. Mean operational and research scores for the screened-out sample were 498 and 387, respectively, with a correlation of 0.42. For the screened-in sample, operational and research means were 454 and 457, respectively, with a correlation of 0.80. All further analyses used only the sample that passed the screening.

Only U. S. citizens are asked to indicate their ethnicity when they complete the biographical questionnaire for the GRE General Test, and not all U. S. citizens supply this information. Ethnicity information was available for 5,146 examinees in the sample for the verbal test. White students comprised the largest group (2,672), followed by Asian-American students (828). The sample also included 246 African-American students, 256 students from the three Hispanic groups (Mexican-American, Puerto Rican, and Other Latino), 93 American Indian students, and 348 examinees who answered "other." In the sample for the quantitative test,

ethnicity information was provided by 4,438 examinees; proportional representation from each group was similar to the sample for the verbal test. The full sample that passed the screening was used for most analyses, with the reduced sample with ethnic identification used for only those analyses that required ethnic information.

Materials

The CAT quantitative and verbal research sections were constituted from regular, full CAT pools (over 300 items each) that did not overlap with pools used for operational CAT sections. The only thing that distinguished the experimental sections from operational sections was timing and the inclusion of a screen indicating that performance on the research section did not contribute to the examinee's official test score. The time limit — extended for some, standard for others — was posted on the beginning screen. As is customary for the GRE examination, a clock showing remaining time was available on screen at all times, though examinees could opt to hide the clock.

Design and Procedures

Each examinee was randomly assigned to one of four experimental groups. The first group was administered the quantitative research test with standard timing (45 minutes), while the second group was administered the quantitative research section with a time limit that was one-and-a-half times the standard allowance (68 minutes). Similarly, the third group was administered the verbal research test with standard timing (30 minutes), and the fourth group was administered the verbal research test with an extended, 45-minute time limit.

Results and Discussion

GRE Quantitative Test

Because prior research on the SAT I (Bridgeman, Trapani, & Curley, 2002) suggests that extra time may differentially affect examinees of different ability levels, the current data were analyzed using the general linear model (GLM), with operational GRE quantitative score treated as a continuous variable. A gender-by-ability-by-timing condition GLM analysis of variance (ANOVA) indicated the expected gender and ability effects and a timing-condition effect (F [1, 6,993] = 40.14, $p < .001$). The three-way interaction and the interaction of timing condition with gender were not significant (Fs < 1), but timing condition did interact with ability (F [1, 6,993] =

23.42, $p < .001$), with a greater effect noted for lower-ability examinees. Although this interaction effect was statistically significant in this very large sample, a regression analysis indicated that it contributed trivially to prediction. The R-square was .807 with gender, ability, and timing condition in the model; adding the ability-by-timing-condition interaction increased R-square by only .001 (i.e., to .808). Table 1 presents the means and standard deviations of the quantitative scores for the two experimental timing conditions.

Table 1

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

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

The interaction, which is ordinal over the entire 200-800 score range, can be seen in the local linear regression lines (with normal kernel smoothing and a bandwidth multiplier of one) shown in Figure 1. With local regression, a point along the smooth curve is derived from a regression of points in the same vicinity, with the closest points receiving heavier weights. This removes some noise in the data, but still allows the regression line to fit any underlying nonlinearity.

Although the interaction is quite small, it is of some interest because it is the exact opposite of the pattern observed in the study of the SAT I: Reasoning test related earlier. The previous study found that extra time on the SAT I did not benefit lower-ability examinees, but did benefit higher-ability examinees (Bridgeman et al., 2002). Here, the minimal effect of extra time on the GRE performance of higher-ability examinees, shown in Figure 1, could simply represent a ceiling effect; examinees who answered nearly all of the questions correctly with standard timing on the operational test had practically no opportunity to demonstrate higher scores with more time. But the presence of a difference in GRE performance, but not SAT I

performance, for lower-ability students who were given more time requires a different explanation.

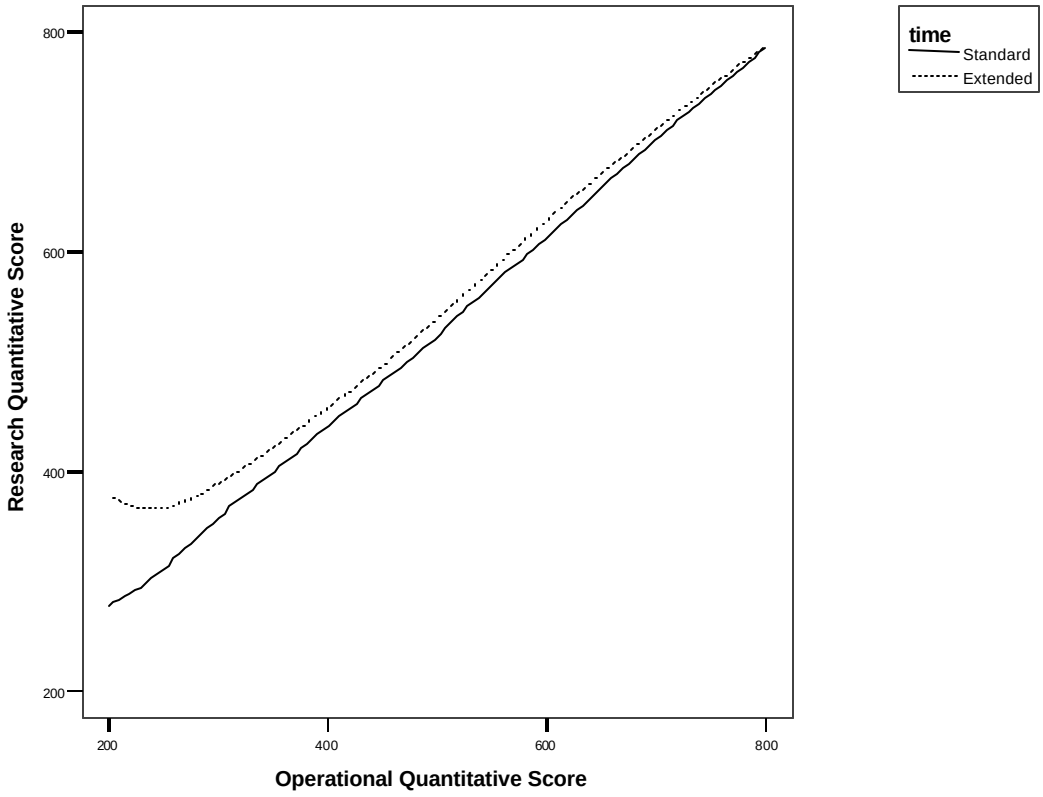

Figure 1. ***Local linear regression of research GRE quantitative score on operational GRE quantitative score for examinees with standard vs. extended time.***

First, it should be noted that the GRE and SAT I populations differ; lower-ability GRE examinees may be more like mid-ability SAT I examinees than lower-ability SAT I examinees. Second, and probably more importantly, the SAT I is a linear test with quantitative items arranged roughly in order of difficulty. Lower-ability examinees who run out of time on the last few, very difficult SAT I items probably would not be able to answer these items correctly even with more time. In contrast, on the computer-adaptive GRE General Test, the last items are targeted to the ability level of the examinee, so that even relatively low-scoring examinees should have a reasonable chance of solving these final items if they have time. In addition, the GRE imposes a penalty for incomplete tests, so it is almost always to the examinee's advantage to complete the test. The SAT I, on the other hand, is a formula-scored test that subtracts a

fraction of a point for a question that is answered incorrectly, but subtracts nothing for questions that are left blank. Thus, lower-ability SAT I examinees with extra time may attempt to answer the difficult questions at the end of the test, be drawn to attractive distracters, and get the items wrong; in the end it is possible that they would have gotten a higher score had they run out of time and left these items blank.

An additional complication with CAT is that examinees at high ability levels are administered more difficult items than examinees of lower ability, and at least for the quantitative and analytical sections of the GRE General Test, evidence suggests that these more difficult items tend to be more time-consuming to complete (Bridgeman & Cline, 2000). This would lead to the expectation that the test is more speeded for higher-ability examinees than lower-ability examinees. Differential speededness for GRE analytical score was indeed observed in another study (Bridgeman & Cline, 2002); very rapid responding on the last few items of the GRE analytical test suggests that higher-ability students were more likely than lower-ability students to run out of time. However, that study did not include quantitative items and did not experimentally manipulate timing; thus, the differential impact of extra time on quantitative items could not be determined. The current results suggest that any differential effects for GRE quantitative score are in the opposite direction (i.e., less speeded for higher-ability examinees than lower-ability examinees); even though higher-ability examinees were administered more time-consuming items, they also benefited less from extra time.

The seven-point difference observed in Table 1 is quite small by conventional standards. The standard deviation of the quantitative research scores in this sample was 123, so the difference is about 0.06 in standard deviation units. Figure 2 depicts these very small differences as a plot of the difference between research and operational scores (research minus operational) for examinees in the standard and extended timing conditions. The discrete points in this graph are the result of a score scale that runs from 200 to 800, but in which the last digit is always a zero. Thus, changes of 10 or 20 points are possible, but a change of 15 points is not possible.

Ethnic differences. A gender-by-ethnicity-by-ability-by-timing condition GLM ANOVA indicated the expected main effects of gender, ethnicity, ability, and timing condition, but no significant two- or three-way interactions with timing condition. Thus, the effects of extra time appear to be relatively constant across ethnicity categories.

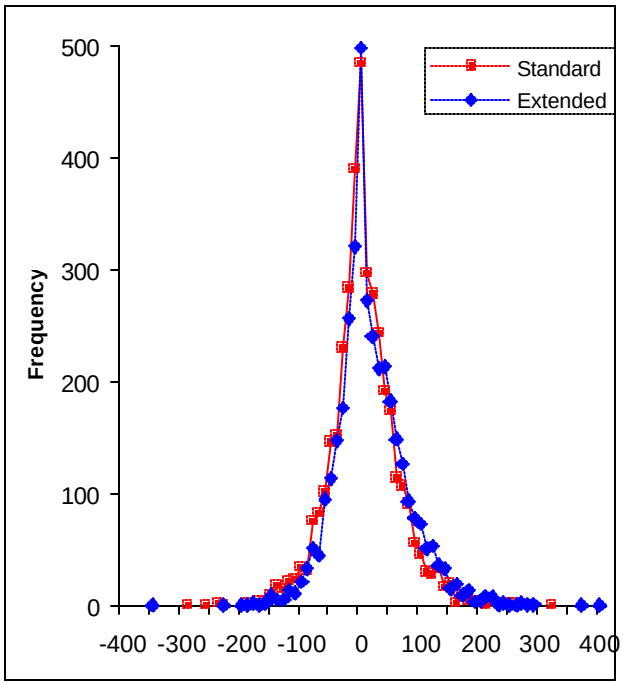

Figure 2. ***Research GRE quantitative score minus operational GRE quantitative score for examinees with standard vs. extended time.***

GRE Verbal Test

The gender-by-ability-by-timing condition GLM ANOVA that was conducted for the research version of the GRE verbal test showed that only ability and gender were statistically significant (alpha = .05). Unlike the results for the research version of the quantitative test, timing condition was not significant (F [1, 7,638]= 1.67, p = .20) for examinees who took the verbal test, nor was the interaction of ability level and timing condition (F < 1). The difference between mean verbal scores for the two timing conditions was seven points — the same difference noted for mean quantitative scores. However, although the sample sizes were comparable, the difference found for the quantitative scores was statistically significant, while the difference between mean verbal scores was not because of a larger mean square error for the latter analysis. (Recall that the correlation of research and operational GRE quantitative scores in the GLM was.90; the same correlation for research and operational verbal scores was .80.) Table 2 shows the means and standard deviations of the verbal scores for the two timing conditions.

Table 2

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

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

Figure 3 displays the local linear regression lines (with normal kernel smoothing) for both timing conditions. The apparent large difference at the very top of the scale may partly reflect fitting a curve for which there are relatively few data points; on the verbal scale, only 42 test takers had scores of 750 or higher, in contrast to the quantitative scale, on which 2,595 examinees scored 750 or higher. Within the range of most of the data (i.e., the 88% of examinees who scored between 300 and 700), the line for the longer test is consistently higher, but by a very small amount.

Figure 4 graphically depicts the differences between research and operational GRE scores for participants in the two timing conditions. As would be expected from the nonsignificant results, the two graphs essentially lie on top of each other.

Ethnic differences. For the reduced sample of examinees who provided ethnic information, the main effect of extra time again was not statistically significant, nor were any of the interactions with extra time significant.

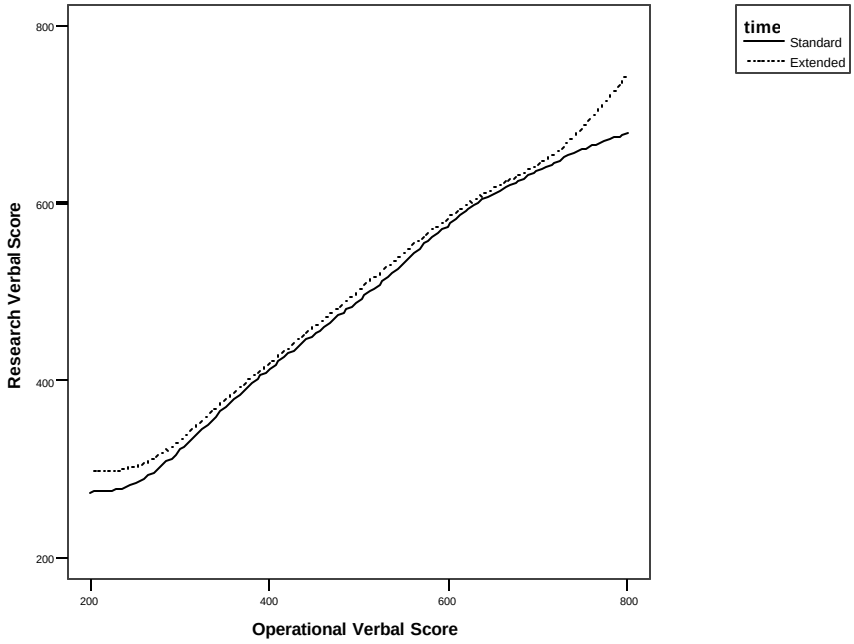

Figure 3. ***Local linear regression of research GRE verbal score on operational GRE verbal score for examinees with standard vs. extended time.***

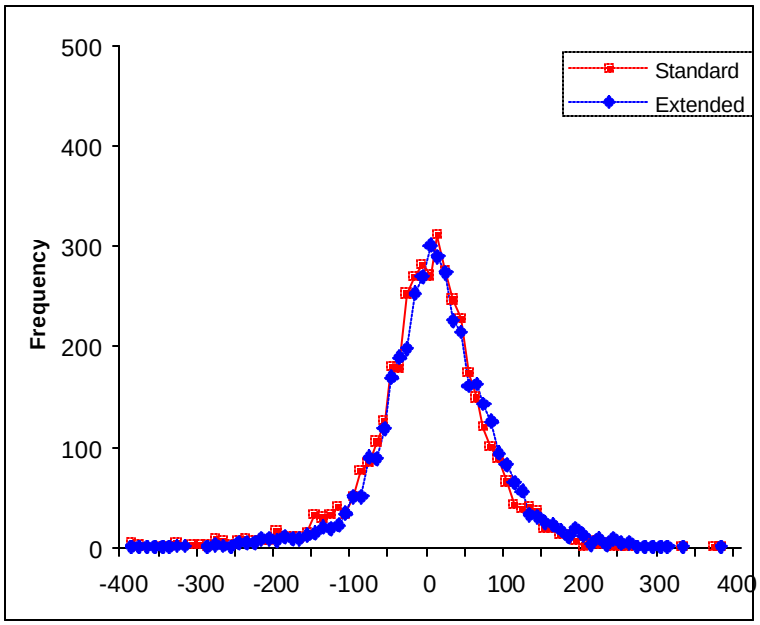

Figure 4. ***Research GRE verbal score minus operational GRE verbal score for examinees with standard vs. extended time.***

Speed-Related Theta Drops

During the computer-adaptive GRE General Test, examinee ability (theta) is estimated after each item is administered. Though these estimates vary substantially early in the test, they tend to become stable toward the end as long as the test is not highly speeded. When the test is so speeded that many examinees resort to guessing on the last five or six items, substantial drops in theta are possible because the scoring algorithm assumes the missed items reflect lower ability, rather than random guessing due to loss of time. In the case of the speededness observed previously for the GRE analytical test, such drops in perceived ability were not infrequent, with about 14% of examinees dropping 0.5 or more theta points over the last six items of the 35-item test (Bridgeman & Cline, 2002). For the quarter of examinees who were under extreme time pressure — defined as having less than two minutes to answer the last six questions — drops of 0.5 theta points or greater were noted for 34% of examinees.

For the current study, we compared changes in theta for examinees in the standard timing condition to changes in theta for examinees in the extended timing condition. For the quantitative test, we studied decreases in theta from item 24 to item 28 (the last item), and for the verbal test we considered changes in theta from item 25 to the last item (item 30). Table 3 shows the percentage of examinees at each theta level for the standard and extended timing conditions on the quantitative test, while Table 4 shows comparable data for the verbal test.

Table 3

Percent of Quantitative Examinees by Timing Condition and Change in Theta

Change in theta (theta at item 24 minus theta at item 28)	Timing condition	
	Standard	Extended
> 0.5	0.4	0.9
-0.5 to 0.5	99.6	99.2
< -0.5	0.0	0.0

Table 4

Percent of Verbal Examinees by Timing Condition and Change in Theta

Change in theta (theta at item 25 minus theta at item 30)	Timing condition	
	Standard	Extended
> 0.5	2.6	2.5
-0.5 to 0.5	96.8	96.8
< -0.5	0.6	0.7

In contrast to the 14% of examinees noted above who dropped 0.5 or more theta points while completing final items of the GRE analytical test (Bridgeman & Cline, 2002), in the current study very few examinees experienced substantial drops in theta over the last few items of the quantitative and verbal tests: less than 1% and less than 3%, respectively. Furthermore, these percentages were virtually identical for both the standard and extended timing conditions, suggesting that the few drops that were noted were not related to time pressures.

Conclusion

Compared to previous findings for the GRE analytical test — or even compared to an absolute standard of no difference between standard and extended timing conditions — the GRE quantitative and verbal tests do not appear to be highly speeded. Allowing one-and-a-half times the standard time would have only a small effect on overall scores and would not differentially affect the performance of women and minorities. However, we cannot rule out the possibility that some groups or individuals not identified in this study could differentially benefit from more test time. If such individuals could be identified, it would be useful to know of any test-taking strategies that might help these students use their time more effectively.

To the extent that the current results show that speed of performance does not appear to be an important part of the reasoning construct assessed by the GRE, these findings are relevant to the debate concerning the provision of accommodations for the disabled . As noted in the professional testing standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the goal of an accommodation is “to minimize the impact of the test taker attributes that are not relevant to

the construct that is the primary focus of the assessment" (p. 101). In addition, the results suggest that test users should not be overly concerned that some students may attempt to gain an unfair score advantage by inappropriately requesting and receiving an accommodation of extra time. Any such advantage would be quite small, especially at the higher score levels where most competitive admissions decisions are made.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (ETS Research Report 85-11). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Cline, F. (2000). *Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assessment* (GRE Board Professional Report No. 96-20P; ETS Research Report 00-7). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Cline, F. (2002). *Effects of differences in expected response times on GRE analytical scores*. Manuscript submitted for publication.

Bridgeman, B., Trapani, C., & Curley, E. (2002). *Impact of fewer questions per section on SAT I scores*. Manuscript submitted for publication.

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

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

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.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimates using the HYBRID model* (TOEFL Technical Report No. 10). Princeton, NJ: Educational Testing Service.