

*Lynch School of Education*

*inTASC Publications*

---

*Boston College*

*Year 2002*

---

Performance Differences According to  
Test Mode and Computer Familiarity on  
a Practice GRE

Amie Goldbert  
Boston College,

Joseph Pedulla  
Boston College,



[www.intasc.org](http://www.intasc.org)

## **Performance Differences According to Test Mode and Computer Familiarity on a Practice GRE**

**Amie Goldberg and Joseph Pedulla**  
**Technology and Assessment Study Collaborative**  
**CSTEEP, Boston College**  
**332 Campion Hall**  
**Chestnut Hill, MA 02467**



## Performance Differences According to Test Mode and Computer Familiarity on a Practice GRE

Amie L. Goldberg and Joseph J. Pedulla  
Technology and Assessment Study Collaborative  
CSTEED, Boston College  
Released April 2002  
IN PRESS at the *Journal of Educational and Psychological Measurement*

### Abstract

Ideally, test performance is unrelated to the mode under which one administers the test. This study investigated relationships between test mode (paper-and-pencil vs. computerized-with-editorial-control, and computerized-without-editorial-control) and computer familiarity (lower, moderate, and higher) with test performance on the Graduate Record Exam (GRE). The GRE test was administered to 222 undergraduate students who were stratified by gender, then randomly assigned to a test mode group. With self-reported GPA as a covariate in a MANCOVA, the authors found that examinees in the paper-and-pencil group outperformed the computerized-without-editorial-control group on all subtests. The computerized-with-editorial-control group outperformed the computerized-without-editorial-control group on the Analytical subtest only. The authors also found a significant main effect for computer familiarity on the Analytical and Quantitative subtests. A significant interaction between computer familiarity and test mode on the Quantitative subtest confounded the main effect for that subtest. The subtests were dramatically more speeded in the computerized forms. Results emphasize the importance of evaluating time constraints when converting exams from paper-and-pencil to computer-delivery.

### Introduction

Moves toward computerized testing stem from the advantages it offers over the traditional paper-and-pencil format. The advantages include: cost-effective administration (fewer materials, proctors, and proctor training), ease, increased accuracy, immediacy of scoring and score reporting, and flexible (even individualized) test scheduling and location. For these reasons, computerized testing now plays an important role in educational and psychological assessment. It should be noted that

there is a down side to computerized testing as well, namely the greater cost of item development because of the increased exposure of items due to an increased number of administrations. These costs can outweigh many of the savings gained through the advantages.

Considerable psychometric research in adapting paper-and-pencil tests to a computer format focuses on mode of administration effects. A meta-analytic study by Mead and Drasgow (1993) compared computerized and paper-and-pencil versions of 123 timed power tests and 36 speeded tests. Their analysis indicated that test mode had no effect on performance for carefully constructed power tests, but a considerable effect for speeded tests. Other research has indicated that specific features of test software are accountable for cross-modal inequivalencies. For instance, multi-screen, graphical, or complex displays have been found to be susceptible to mode effects (Mazzeo & Harvey, 1988). Additionally, scrollable text, lack of work space (Lee & Hopkins, 1985), and the inability to review and revise responses (Wise & Plake, 1989; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992) have also been identified as features that can significantly affect examinee performance.

Item review, i.e. the examinee's ability to return to previous items, is actually part of a larger issue concerning the amount of control that examinees should have in computer based testing (CBT). Wise & Plake (1989) noted three basic features regarding item review potentially available to examinees during paper-and-pencil tests that should be considered by CBT developers: 1. allowing examinees to skip items and answer them later in the test, 2. allowing examinees to review items already answered, and 3. allowing examinees to change answers to items.

In 1992, the Educational Testing Service (ETS) began delivering the Graduate Record Exam (GRE) General Test in computerized form in addition to the traditional paper-and-pencil form. ETS's transition towards a completely computerized GRE test program spanned seven years, ending in the Spring of 1999. Converting the GRE to a computer-adaptive test (CAT) involved two major changes, and it was apparent to ETS that each of these changes could potentially affect the relationship between the paper-and-pencil and CAT GRE General Test. Therefore, in 1991, ETS employed a two-stage investigation of transitioning from paper-and-pencil to CAT. The first step compared the linear (fixed-item) paper-and-pencil test to a linear computer based test (CBT). The second step compared the linear CBT to a CAT. The CAT is a non-linear CBT that employs item response theory (IRT) methods to create a tailored test for each examinee. In the adaptive test paradigm, item selection is based on examinee ability, which is re-estimated after each recorded item response.

Based on the findings from the first investigation, the CBT GRE was deemed 'comparable' with the paper-and-pencil version. Specifically, ETS concluded that the Verbal and Analytical subtests in particular did not demonstrate any differences between test modes. And, although a "slight test-level mode effect" was found for the Quantitative scale, "no adjustment was made" (Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995, p. 29).

Limitations in the design and interpretation of findings from ETS's first study exist. The field test sample was recruited from a national GRE administration and was re-administered a CBT GRE at varying times in the following two to nine weeks. ETS acknowledged that it was not possible to completely remove or disentangle any retest

effects that may be present. Also, ETS's CBT, akin to the paper-and-pencil version, provided examinees with editorial control (the ability to view, respond to, skip, and review subtest items) on six of seven test sections. Only for the last section, a Verbal subtest, was the effect of suppressing this feature investigated.

In March 1993, Phase Two of ETS's CAT research began. ETS compared the CBT with a CAT. The CAT's first six of seven sections were actually in a linear CBT format and included a "review" option. ETS also designed a CAT version for each subtest (Analytical, Quantitative, and Verbal), and each examinee received one of these three different CAT's as his/her seventh section. The purpose of administering this seventh section was to verify that the scores derived from the CAT measure had similar characteristics to scores derived from the linear CBT, and thus by inference, similar to those for the paper-and-pencil form.

ETS found that both computerized formats yielded "adequately comparable scores" for the Verbal and Quantitative test sections, "although the Analytical section provided the most mixed results" (Schaeffer et al., 1995, p. 35). After statistically adjusting the Analytical scale, ETS implemented its computer adaptive GRE test program.

### Purpose of the Study

The present study is designed to investigate mode effects between paper-and-pencil and computer-based linear formats on GRE General Test performance, which was the focus of the first phase of ETS's field test of the CBT GRE. The design of the present study differs from that of ETS's original design in several ways. As detailed in the methodology section, this study's design incorporates a full analysis of suppressing and enabling editorial control on all three GRE subtests, and includes analyses based on level of examinee computer familiarity. This design excludes the confounded retest and mode effects that are present in the original field trial of the computerized GRE. Also, in contrast to the original study, no incentives for participation were offered.

## Methods

### Instrumentation

ETS's Permissions Department granted permission to use sections four, five, and six of the GRE General Test GR86-2, published in *Practicing to Take the GRE General Test, 9th Edition* (GRE Board, 1992, pp. 182–200), though in no way does ETS endorse or have any connection with this study. This modified GRE, which included one section per subtest (Analytical, Verbal, and Quantitative), was administered under three different test conditions. Although it is possible that participants may have had previous exposure to these specific practice items, it should be noted that a practice version older than the most recently released one was used so that possible familiarity with the practice test by the examinees in this study would be minimized.

In the first condition, the modified GRE was administered in a paper-and-pencil format, closely resembling that of the traditional GRE General Test administration. Efforts were made to duplicate the GRE standardized test environment. Group

instructions, time constraints, machine scannable answer sheets, number two pencils, and scratch paper were included.

The second and third test conditions, although computerized, contained the same items as the paper-and-pencil condition. The first form of the computer-based fixed-item tests was modeled after the original CBT GRE, pilot tested in the Fall of 1991 by ETS (Schaeffer et al., 1993). In this condition, examinees were provided with a clock on the computer screen (with the option of viewing or not viewing the clock), as well as editorial control in responding to items on all three GRE subtests, i.e. the ability to change responses or skip items and come back to them later. In the third condition, the second computerized form, the examinees were provided with a clock (with the option of viewing or not viewing the clock), but no editorial control on any of the subtests. These two computerized conditions enabled the investigation of performance differences attributable to examinee control over ability to edit responses.

Several test items contained in the Verbal and Quantitative subtests required examinees in the computerized test groups to return to a previous screen to reference a reading passage or graphic. To maintain as much consistency as possible among the three groups, reading passages and graphs were displayed on separate pages from their corresponding items to the paper-and-pencil test group also.

Directions for the computerized GRE forms were shown on the screen prior to the start of the test. The test clock did not begin until the screen containing the first item was displayed. Directions for both of the computerized versions differed from the standard GRE directions in several ways. All administrations began with verbal directions by the proctor. The examinees in the computerized groups were able to follow along with these instructions as they were displayed on their screen; the examinees in the paper-and-pencil group were able to follow along with these instructions in their test booklet. Both computer groups were advised that they could use scratch paper and pencils (provided by the proctor) to aid them in answering any item on the test. During the instructions to the computerized groups, two sample items were provided so that all examinees could practice how to record item responses. After the examinees practiced with these two items, the test proctor answered any questions related to the answer recording process.

Both computerized versions employed a two step procedure for finalizing the recording of an answer. After indicating a response, examinees were asked to verify that response a second time before it would be recorded as their answer. The computerized test instructions differed between the two versions in one way: examinees in the computer group with editorial control received directions on how to use the item menu bar to move among items within each test section; directions for examinees without editorial control informed them that they had to respond to the items in sequence and could not change a response once they finalized it. As with ETS's current computerized GRE format, both computerized exams presented just one item per screen.

The GRE subtests administered in this study were originally administered by ETS in paper-and-pencil format only, where scaled scores would be calculated using two sections per subtest. For example, a Verbal GRE scaled score is a score based on two Verbal sections. Correct answers are summed, and the total number of correct responses on each subtest (across both sections) corresponds to a particular scaled

score. To ease the burden and decrease any fatigue effect on the examinee, all three conditions of this study included just one section from each of the three GRE General Test domains. Therefore, the data analyzed in this study's analyses are the raw scores on each of these test sections. The total number of items in the Analytical, Verbal, and Quantitative sections were 25, 38, and 30, respectively. The original GRE field test increased testing time by two seconds per item for each section to account for the time used by the computer program for "screen refreshing" (Schaeffer et al., 1993). Similarly, but more liberally, for the present study there was a five second adjustment per item added to the allotted testing time for each subtest.

### Sample

Two hundred and twenty-two students from two colleges on the East coast volunteered their participation. Though participants were first stratified by gender and then randomly assigned to a test condition, the extreme imbalance of male ( $n=63$ ) and female ( $n=159$ ) participants did not allow for meaningful analyses involving gender. All students were traditional-aged, third- or fourth-year undergraduates. Characteristics of this study's sample limit the extent to which the results are generalizable to the typical GRE test-taking population. Specifically, this sample deviates from the larger GRE test-taking population in the following ways. This sample consists solely of traditional aged college students, whereas the general GRE test taking population also includes a segment of test-takers who graduated college some time ago. Also, in terms of gender representation, this sample is considerably gender imbalanced; there are approximately two and half times as many females as males. This imbalance is not reflected in the actual GRE test-taking population. The students in this sample took the GRE as a practice test. Whether they were motivated to do their best work is unknown. Finally, in terms of educational background, all of the participants in this study attended one of two medium-sized selective liberal arts colleges from the Northeast, whereas the actual GRE test-taking population covers a much broader geographic and academic range. These differences between the study's sample and the GRE need to be considered when interpreting this study's results.

### Mean Performance Analyses

This study addressed the following three research questions:

- 1) Are there differences in mean performance among CBT (with and without editorial control) and paper-and-pencil versions of each subtest of the GRE General Test?
- 2) What is the relationship between computer familiarity and CBT score?
- 3) Is there a significant interaction effect between test mode and computer familiarity on CBT score?

Descriptive statistics and a multivariate analysis of covariance (MANCOVA) were used to examine the relationship between test condition, computer familiarity, and the interaction between test condition and computer familiarity on the three GRE subtest scores simultaneously, while adjusting for examinees' self-reported SAT and GPA (covariates).

## Test Speededness Analyses

The GRE has characteristics of both a power and speeded test. The GRE is a power test in terms of content and is administered under standardized timed conditions. Historically, much evidence has been presented which suggests that response speed and response power are best regarded as different attributes (Terranova, 1972). Ideally, test speededness could be analyzed by comparing the test performance of examinees who have been administered parallel forms of a test under both timed and untimed test conditions. If the test was speeded, we would find that the additional time provided to examinees increases their test performance.

According to Swineford (1973) and Schaeffer et al. (1993), ETS regards a test as possibly speeded when (1) fewer than 100% of the examinees reach 75% of the test, (2) fewer than 100% of the items are reached by 80% of the examinees, and/or (3) when the ratio of “not-reached variance” to “total score variance” is greater than 0.15. Though these criteria are somewhat arbitrary, ETS maintains that they provide preliminary measures of test-speededness. Donlon (1979) noted that when a test meets any of these preliminary criteria, it is likely that the test is speeded. These three indices of test-speededness, in addition to “percentage of sample answering last item,” were employed in ETS’s analysis of the computerized GRE field test data and will be used for comparisons among the three test conditions in the present study. These four perspectives of test speededness will provide the data necessary to address the fourth research question: are there differences in measures of test speededness among the CBT’s (with and without editorial control) and paper-and-pencil versions of each subtest of the GRE General Test, and what is the relationship between test scores and these measures of speededness?

## Results

A demographic profile of the sample for the present study and an assessment of how much deviation exists between it and ETS’s field test sample serve as a useful backdrop for the later analyses that examine performance on each GRE subtest according to test condition and computer familiarity.

This study’s sample ( $n=222$ ) includes 159 females (71.6%) and 63 males (28.4%). The entire sample consists of traditional-aged undergraduate students. The vast majority (82%, of the participants) identify themselves as White, 5.0% as Asian-American/Pacific Islander, 5.0% as Other, 4.1% as African-American/Black, and 3.6% as Hispanic/Latin-American. The demographic profiles of each test mode group are generally consistent with that of the overall group. In the present study, there are considerably more females than in the ETS field test of the computerized GRE (72% vs. 58%). In terms of ethnicity, Whites constitute the overwhelming majority of participants, while each of the other ethnic groups account for less than 10% of the participants in each study.

## Sample Classification by Independent Variables and Gender

The derivation of the computer familiarity scale was based on the sum of item responses to a series of questions dealing with the participants’ familiarity with specific computer hardware, software, and the frequency with which they used various computer

skills. The total number of possible points on this scale ranged from 0 to 31; the scores from the sample in the present study ranged from 19 to 31. Participants were then categorized into three groups: the lower computer familiarity group consisted of participants scoring below 23 ( $n=27$ , 12.4%); the moderate computer familiarity group consisted of those scoring in the 23 to 27 range inclusive ( $n=128$ , 58.7%); while participants scoring above 27 constituted the higher computer familiarity group ( $n=64$ , 29.4%). Three participants did not answer the computer familiarity items and are not included in the analyses that use this variable. Although the numbers in many of the subgroups were small (particularly for males and the lower computer familiarity group—in a few instances 4 or 5 cases, often 10 to 20), the relative proportions were equivalent across groups.

### Multivariate Analysis of Covariance of Subtest Performance by Test Mode, Computer Familiarity and the Interaction Between Test Mode and Computer Familiarity

MANCOVA was employed to incorporate all three dependent variables (Verbal, Quantitative, and Analytical raw scores) in a single analysis. The independent variables were test mode group, computer familiarity, and the interaction between test mode group and computer familiarity, with undergraduate GPA and SAT scores used as covariates. A test of multivariate normality indicated that SAT interacted significantly with test mode; SAT was therefore dropped from the analysis.

The multivariate tests, using Wilks' Lambda, showed significant main effects for test mode ( $p < .001$ ) and computer familiarity ( $p < .004$ ) (see Table 1). Following the multivariate test, the tests of between subject effects indicated performance significantly differed among test modes on each of the subtests: Analytical ( $p < .002$ ); Verbal ( $p < .002$ ), and Quantitative ( $p < .001$ ). Significant main effects were also found for computer familiarity on the Quantitative ( $p < .001$ ) and Analytical ( $p < .004$ ) subtests. Finally, there was a significant interaction effect between test mode and computer familiarity on Quantitative subtest performance ( $p < .042$ ).

**Table 1: Multivariate Test of the Main and Interaction Effects on Verbal, Quantitative, and Analytical GRE Raw Subtest Scores**

Source of Variance	Wilks' Lambda	F	Hypothesis df	Error df	Sig.
<b>GPA</b>	.897	7.852	3	206	.001
<b>Test Mode</b>	.859	5.443	6	412	.001
<b>Computer Familiarity</b>	.908	3.379	6	412	.004
<b>Test Mode x Computer Familiarity</b>	.930	1.261	12	545	.238

Post-hoc comparisons for each subtest by test mode group were performed applying the Bonferroni correction. Table 2, which contains the mean total raw scores for the Analytical, Verbal, and Quantitative subtests, adjusted for undergraduate GPA, across the three tests groups, clarifies the findings of the post-hoc comparisons. On all three subtests, the paper-and-pencil group significantly outperformed the computerized without editorial control group. Specifically, the paper-and-pencil group outperformed the computerized without editorial control group by 2.8 (of a possible total of 25) raw score points on the Analytical subtest ( $p < .003$ ), 2.6 (of a possible total of 38) raw score points on the Verbal subtest ( $p < .022$ ), and 3.0 (of a possible total of 30) raw score points on the Quantitative subtest ( $p < .001$ ).

**Table 2: Mean Raw Scores by Test Mode for Analytical, Verbal, and Quantitative Subtests Adjusted for Undergraduate GPA**

Subtest	Test Mode	Mean	Standard Deviation
Analytical	Paper-and-Pencil	17.4	4.84
	Computerized without Editorial Control	14.6	5.17
	Computerized with Editorial Control	16.3	6.11
Verbal	Paper-and-Pencil	24.1	5.54
	Computerized without Editorial Control	21.6	5.91
	Computerized with Editorial Control	22.0	6.99
Quantitative	Paper-and-Pencil	22.3	4.42
	Computerized without Editorial Control	19.3	4.71
	Computerized with Editorial Control	20.3	5.57

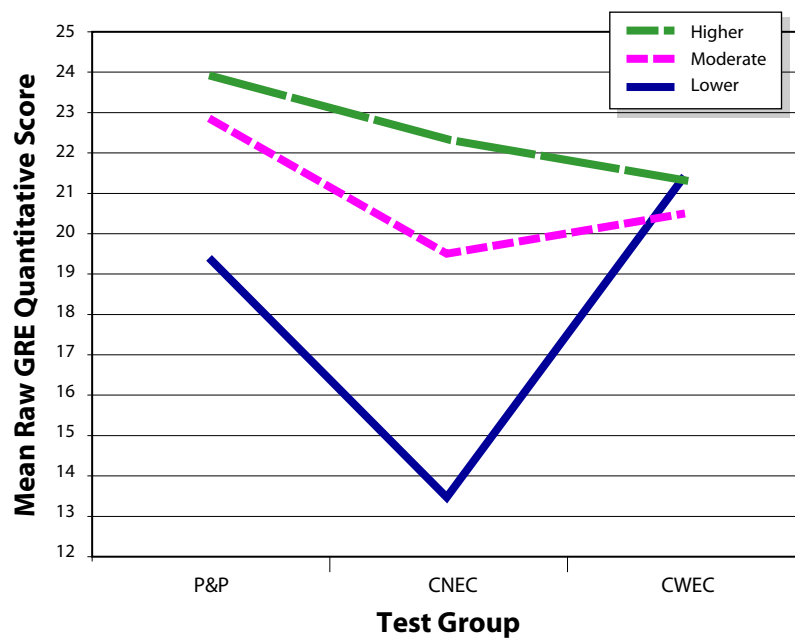
Post-hoc analyses on scaled scores adjusted for undergraduate GPA (see Table 3) indicate that the higher computer familiarity group significantly outperformed the lower computer familiarity group by 2.70 raw score points on the Analytical subtest ( $p < .03$ ), and outperformed the moderate and lower computer familiarity groups by 3.2 ( $p < .003$ ) and 1.50 raw score points ( $p < .042$ ) respectively) on the Quantitative subtest. This finding must be interpreted in light of the significant interaction between computer familiarity and test mode for the Quantitative subtest. Figure 1 shows that

within the computerized with editorial control condition, all three familiarity groups perform equally well. The lower familiarity group scores lower than the other two groups under the paper-and-pencil condition and much lower under the computerized without editorial control condition.

**Table 3: Mean Raw Scores, by Level of Computer Familiarity for Analytical, Verbal, and Quantitative Subtests Adjusted for Undergraduate GPA**

Subtest	Computer Familiarity	Mean	Standard Deviation
Analytical	Lower	14.8	4.58
	Moderate	16.1	4.28
	Higher	17.5	4.33
Verbal	Lower	22.3	5.24
	Moderate	22.5	4.90
	Higher	22.9	4.94
Quantitative	Lower	19.0	4.18
	Moderate	20.7	3.90
	Higher	22.2	3.94

**Figure 1: Interaction Effect Between Test Mode and Computer Familiarity on Mean GRE Quantitative Score**



### Effect Size Analysis

Another way to look at these differences among groups is to consider their magnitude through the measure of effect size. Effect size, as expressed by Cohen's  $d$ , describes the difference between the mean of an experimental group (in this case, both computerized test groups) and a control group (the paper-and-pencil group), in standard deviation units. In other words, effect size is simply the  $z$ -score of the mean of the experimental group referenced in the frequency distribution of the control group (Glass & Hopkins, 1996, p. 290).

The effect sizes and respective confidence intervals for test mode are contained in Table 4. Hedges & Olkin's (1985) formula for calculating confidence intervals around effect sizes was employed:

$$\sigma[d] = \sqrt{\frac{N_E + N_C}{N_E \times N_C} + \frac{d^2}{2(N_E + N_C)}}$$

where  $N_E$  and  $N_C$  are the numbers in the experimental and control groups, respectively. Following that, the 95% confidence interval for  $d$  would be:

$$d - 1.96 \times \sigma[d] \text{ to } d + 1.96 \times \sigma[d]$$

Cohen's (1977) guidelines for interpreting effect sizes suggest that effect sizes (in absolute value) of approximately .25 are considered small, .50 medium, and .75 large. According to these guidelines, the effect sizes for the differences between the paper-and-pencil and computerized without editorial control groups for all three subtests fall in the medium to upper-medium range. The effect sizes for the differences between the paper-and-pencil and computerized with editorial control groups for the Verbal and Quantitative subtests are in the lower medium range; for the Analytical subtest in the small range (in fact this confidence interval contains 0).

**Table 4: Effect Sizes with Confidence Intervals for Test Mode on Analytical, Verbal, and Quantitative Subtests**

Difference Between Groups		Analytical Subtest	Verbal Subtest	Quantitative Subtest
Paper-and-Pencil and Computerized <b>without</b> Editorial Control	Effect Size	-.587	-.462	-.679
	C.I.	-.838, -.287	-.717, -.170	-.929, -.374
Paper-and-Pencil and Computerized <b>with</b> Editorial Control	Effect Size	-.225	-.375	-.437
	C.I.	-.477, .084	-.610, -.046	-.664, -.099

### Measures of Test Speededness by Test Mode

Four measures were used to summarize the degree of test speededness for each subtest by test mode: 1. percent reaching last item, 2. percent reaching 75% of items, 3. number of items reached by 80% of the examinees, and 4. ratio of “not reached” variance to “total score” variance. “Not reached variance” represents the variance of the number of items left unanswered following the last item for which a participant responded. This statistic was divided by the “total score” variance in order to obtain the “not reached” to “total score” variance ratio. As shown in Table 5, the paper-and-pencil test was the least speeded of the three test modes for all three subtests across all four speededness measures. The most striking differences among test mode groups were found in “Percent Reaching Last Item” and in the variance ratios. The differences among test groups in the percentage of participants reaching the last item on the Analytical subtest was quite large (77% for the paper-and-pencil, 31% for computer with no editorial control, and 39% for computer with editorial control).

Using a ratio of greater than or equal to 0.15 as the criterion indicative of test speededness, the variance ratios among the three test modes are different as well. Table 5 shows that the Analytical subtest was highly speeded in all three test modes, but to the greatest extent in the computerized with no editorial control group. Both the Verbal and Analytical subtests were speeded in both computerized test modes, and again, most markedly in the computerized with no editorial control mode. These two subtests were not speeded for the paper-and-pencil group. The number of Verbal items reached by 80% of the sample in the computerized without editorial control group was less than that reached by the other two groups, especially by the paper-and-pencil group. An examination of the correlations between raw subtest score and the ratio of “not reached” variance to “total score” variance ratios indicates significant, negative relationships between test speededness and test performance for each of the three subtests. Generally, these correlations were small (in the -.1 to -.3 range). These results, in conjunction with those above, suggest that time constraints negatively affected test performance in both computerized modes.

**Table 5: Measures of Test Speededness on the Verbal, Quantitative, and Analytical Subtests by Test Mode**

Subtest	Percent Reaching Last Item			Percent Reaching 75% of items			Number of items reached by 80% of the sample			Not Reached Variance to Total Score Ratio		
	Paper-and-Pencil	Computerized Without Editorial Control	Computerized With Editorial Control	Paper-and-Pencil	Computerized Without Editorial Control	Computerized With Editorial Control	Paper-and-Pencil	Computerized Without Editorial Control	Computerized With Editorial Control	Paper-and-Pencil	Computerized Without Editorial Control	Computerized With Editorial Control
<b>Verbal</b> (n items = 38)	95.7	77.8	59.3	98.6	92.6	91.5	38	26	33	0.12	0.48	0.41
<b>Quantitative</b> (n items = 30)	88.6	69.1	63.4	100	93.8	97.2	29	28	28	0.15	0.36	0.34
<b>Analytical</b> (n items = 25)	77.1	30.9	39.4	87.1	79.0	85.9	20	18	19	0.45	0.64	0.47

## Discussion

Though this study's sample profile departs from the GRE General Test population (especially in regard to gender), significant performance differences were found among test modes on a practice GRE. These performance differences are in contrast to similarities found by ETS in their field study that compared the paper-and-pencil GRE with a computer-based GRE.

MANCOVA revealed test mode had a significant main effect on test performance on all three GRE General subtests, and level of computer familiarity was related to the Quantitative and Analytical subtests. When controlling for ability as measured by undergraduate GPA, participants in the paper-and-pencil group significantly outperformed participants in the computerized without editorial control group on each of the subtests by approximately 2.5 to 3.0 raw score points. These findings suggest that converting a test for computer delivery requires additional considerations, if the computerized test scores are to be used interchangeably with traditional paper-and-pencil scores.

Test-speededness analyses provided insight into these differences (and lack thereof, when comparing the two computerized test groups) in mean performance. Though the Analytical subtest was markedly speeded across all three of the test modes, test speededness was most dramatic on the Verbal and Quantitative subtests in both computerized test modes, and most speeded in the computerized without editorial control test mode on all three subtests. Therefore, although no significant differences in mean performance were found between the paper-and-pencil and computerized with editorial control groups, these two test modes provided different test-taking

experiences, in terms of speededness. Findings from this analysis highlight the fact that the two to three extra minutes included in each computerized subtest administration for ‘screen refreshing’ (5 seconds per item) was not an adequate time adjustment. Consequently, it is likely that those participants who were provided editorial control in each subtest did not have adequate time to use this test feature effectively. Therefore, even greater increases in time allotted in order to establish cross-modal test equivalency is a worthy consideration.

Directions for future research in this area include revisiting the editorial control feature in computerized tests under more liberal time allocations. Due to the increased test-speededness found in this study, the effect of enabling and suppressing editorial control on test performance was not fully addressed. The effect of computer familiarity on CBT performance also deserves further attention, especially considering the proliferation of computerized tests beyond those designed primarily for student populations such as in this study. In particular, it is likely that computer usage is more variable among people who are not part of the traditional undergraduate age group and among people who are more socioeconomically diverse. The increased computerization of such measures as clinical psychological tests and personnel selection tests (vocational certification, tests of cognitive ability, etc.) demands research that is designed to focus on demographically diverse populations. Such research is the essential prerequisite for making valid and responsible decisions based on cross-modal test score interpretations.



## References

- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Donlon, T. F. (1979). *Time for review as an aspect of test speededness*. Paper presented at the annual meeting of the New England Educational Research Organization (Provincetown, MA, April 29–May 1, 1979). ERIC Document Reproduction Service No. ED 181 034.
- Glass, G.V. & Hopkins, K.D. (1996). *Statistical methods in education and psychology*, 3rd ed. Boston: Allyn and Bacon.
- Graduate Record Examinations Board (1992). *Practicing to take the GRE general test, 9th ed.* Princeton, NJ : Educational Testing Service.
- Hedges, L. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*. New York: Academic Press.
- Lee, J.A., & Hopkins, L. (1985). *The Effects of training on computerized aptitude test performance and anxiety*. Paper presented at the annual meeting of the Eastern Psychological Association (56th, Boston, MA, March 21–24, 1985) ERIC Document Reproduction Service No. ED 263 889.
- Mazzeo, J., & Harvey, A.L. (1988). *The Equivalence of scores from automated and conventional educational and psychological tests: a review of the literature*. (ETS Report No. RR-88-21) Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 304 462).
- Mead, A.D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin*, 114(3), 449–58.
- Schaeffer, G.A., Reese, C.M., Steffen, M., McKinley, R.L., & Mills, C.N. (1993). *Field test of a computer-based GRE General Test* (ETS Report No. RR-93-07). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 385 588).
- Schaeffer, G.A., Steffen, M., Golub-Smith, M.L., Mills, C.N., & Durso, R. (1995). *The Introduction and comparability of the computer adaptive GRE General Test* (ETS Report No. RR-95-20). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 393 936).
- Swineford, F. (1973). *An Assessment of the Kuder-Richardson Formula (20) reliability estimate for moderately speeded tests*. Paper presented at the annual meeting of the National Council of Measurement in Education (New Orleans, Louisiana, February 28, 1973). ERIC Document Reproduction Service No. ED 074 095.
- Terranova, C. (1972). Relationship between test score and test time. *Journal of Experimental Education*, 40(3), 81-3.
- Vispoel, W.P., Wang, T., de la Torre, R., Bleiler, T., & Dings J. (1992). *How review options and administration modes influence scores on computerized vocabulary tests*. Paper presented at the annual meeting of the American Educational Research Association (San Francisco, CA, April 20–24, 1992) ERIC Document Reproduction Service No. ED 346 161.

Wise, S.L., & Plake, B.S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 8(3), 5–10.

