DOCUMENT RESUME

ED 401 312 TM 025 880

AUTHOR	Slater, Sharon C.; Schaeffer, Gary A.
TITLE	Computing Scores for Incomplete GRE General Computer Adaptive Tests.
PUB DATE	Apr 96
NOTE	36p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; College Students; *Computer Assisted Testing; Equal Education; Higher Education; *Psychometrics; Scores; *Scoring; Simulation; Test Bias; *Test Results; *Timed Tests
IDENTIFIERS	*Graduate Record Examinations; Incomplete Data Sets; Monitoring

ABSTRACT

The General Computer Adaptive Test (CAT) of the Graduate Record Examinations (GRE) includes three operational sections that are separately timed and scored. A "no score" is reported if the examinee answers fewer than 80% of the items or if the examinee does not answer all of the items and leaves the section before time expires. The 80% threshold was adopted to set a minimum threshold to result in psychometrically acceptable scores without penalizing slow test takers. A study was conducted to examine the impact of possible CAT test-taking (and test completion) strategies and scoring options with regard to incomplete tests and to consider the impact of working to the 80% level or beyond it. Actual data for different completion points were selected from 70,000 GRE CAT examinees. CAT simulations were also conducted to assess the impact of different test taking strategies and different ways of scoring tests when not all items were answered. Data did not indicate that there is widespread use of the 80% rule as a strategy for deciding how much of the GRE CAT to complete. Serious psychometric and equity issues would be raised if examinees began to employ this strategy, so monitoring of examinee strategies is important. (Contains three figures and eight tables.) (SLD)

COMPUTING SCORES FOR INCOMPLETE GRE GENERAL COMPUTER ADAPTIVE TESTS*

Sharon C. Slater, University of Massachusetts, Amherst

Gary A. Schaeffer, Educational Testing Service

Paper presented as part of the symposium "CATucopia:
Measurement Issues Faced by a Large-Scale Testing Program"
at the annual meeting of the
National Council on Measurement in Education,
New York, New York, April 1996.

* This project was conducted while Sharon Slater was a summer intern at Educational Testing Service in 1995. The authors acknowledge Craig Mills of ETS for initiating the project. Marna Golub-Smith, Craig Mills, Judson Sheridan, and Manfred Steffen, all of ETS, provided thoughtful reviews of drafts of this paper. The authors also acknowledge the technical expertise offered by Manfred Steffen and the data processing assistance provided by Tama Braswell and Marion Horta. Points of view or opinions stated in this report do not necessarily represent official Graduate Record Examination Board position or policy.

Computing Scores for Incomplete GRE General Computer Adaptive Tests

Introduction

The GRE General computer adaptive test (CAT) includes three operational sections that are separately timed and scored. The GRE CAT score reporting policy for the verbal, quantitative, and analytical sections can be summarized as follows:

1. A CAT score for a section is reported if a) the examinee answers all items in the section, or b) the examinee answers at least 80% of the items in the section and section time expires. The score is based on the ability estimate computed after the last item answered.

2. A No Score is reported if a) the examinee answers fewer than 80% of the items in the section, or b) the examinee does not answer all of the items and leaves the section before time expires.

There were a number of reasons why the 80% threshold was adopted as opposed to either requiring examinees to answer all items or allowing examinees to answer as few as one item to get a score. The initial motivation for not requiring examinees to answer all items was due to equity. It was believed that examinees who work somewhat more slowly than others (and therefore would not finish the test in the allocated time) should be allowed to receive a score. However, some minimum threshold was required because with maximum likelihood scoring (like that in the GRE CAT), it would otherwise be possible to get a very high score by answering as few as one item correct and not answering any other items. Such a score, of course, would raise insurmountable reliability and validity issues and would never be reported. Therefore, a threshold was needed that would result in psychometrically acceptable scores while not penalizing slow test takers. The final threshold of 80% was selected because simulation results indicated that 80% length CATs produced scores that were adequately precise in terms of reliability, conditional standard errors of measurement, and content representativeness. In addition, based on linear CBT timing data, it was also believed that almost all examinees would be able to answer all of the items in the specified time limits.

The 80% threshold, which has been in effect operationally since November 1993, has raised some concerns. Two of these concerns are a) the potential use of the 80% threshold as a test-taking strategy for maximizing scores, and b) not reporting scores for examinees who answer fewer than 80% of the items. Regarding the first concern, it may be possible for examinees to use the threshold to their advantage to produce a positive bias in their estimate of ability. In fact, the question arises as to why examinees would choose to answer items beyond the threshold. For example, examinees could pace themselves in order to devote the maximum amount of time to answering the minimum number of items required to receive a score. This would maximize the average time spent per item. This may result in a positive bias in examinees' scores, that is, examinees may receive higher scores if they pace themselves to answer only 80% of the items than if they had spent less time per item and answered all items. On the other hand, this strategy would result in these examinees being administered harder items, and the additional time per item may not be sufficient for them to correctly answer these questions. That is, to some degree the CAT may self-correct.

The threshold also can have serious implications for test-takers who receive a 'No Score'. These examinees may be passed over for admissions or fellowships because they do not have one or more GRE scores reported. Or, these examinees may be required to repeat the test if a score is required for admissions or a fellowship application.

The purpose of this study was to investigate the impact of possible CAT test-taking strategies and scoring options with regard to incomplete tests. There is no single best test-taking strategy for every examinee because examinees do not know the content or difficulty level of subsequent items as they proceed through the CAT. However, there are some general strategies that examinees may choose to implement. The test-taking strategies examined in this study include the following:

1. *Answer only the minimum number of items required to receive a score.* Examinees may pace themselves in order to maximize the time spent per item. This assumes that the probability of a correct response increases as time spent on an item increases.

2. *After the minimum number of items have been answered, continue to answer items until the probability of a correct response to the next item is perceived to be low. Then, do not answer that item and let time expire.* This strategy may increase the likelihood of maximizing an examinee's score.

3. *Answer all items.* This is the strategy that most examinees are accustomed to following.

Both actual and simulated CAT data were used to address these strategies.

In an ideal psychometric world, all examinees would have the time and motivation to answer all items. This would allow for the most accurate assessment of the abilities being measured by the test. However, in the real world many examinees do not answer all items. One way of trying to encourage examinees to answer all items is to employ scoring rules for incomplete CATs that would encourage examinees to complete the test. These scoring rules would have the purpose of minimizing or negating the benefits of employing a strategy to use the 80% rule to their advantage. They also may penalize examinees who truly work slowly, an outcome that differs from the intent in establishing the original 80% rule in the first place.

Two scoring rules for incomplete CATs were explored in this study using simulated data. The purpose of these rules is to lower the scores of examinees who do not finish their test below what they would receive if the score were based only on the items answered. Because these rules would be publicized, they would also encourage examinees to complete their test. In one scoring rule, random right/wrong responses were assigned to unanswered items. In the other scoring rule, wrong responses were assigned to unanswered items. The first scoring rule mimics what P&P examinees may do when section time is about to expire. The second scoring rule is the same penalty function used in the P&P program. These scoring rules were applied to simulated cases where responses were imputed from the 80% point to the end. This allowed for comparing scores at the 80% point with scores from the 'completed' CAT.

Preview of Remainder of Report

The next section will examine data from an analysis sample of actual CAT examinees. These data will be used to explore how examinees proceed through the CAT in terms of numbers of items answered, times spent on each section, and distributions of item times. In addition, actual data will be used to address the issue of using the 80% rule as a test-taking strategy by looking at the differences in scores between 80% and 100% for examinees who completed their CAT.

Following that, simulation data will be used to address the potential impact of the test taking strategies of answering only 80% of the items and of continuing beyond the 80% point until the answer to question is not known. Simulated data then addressed the potential impact of scoring rules where unanswered items are marked either as wrong or are marked at random. The final section will be a discussion of the results.

Analysis Sample for Actual CAT Data

The study used data from approximately 70,000 actual GRE CAT examinees. The actual CAT data were from operational administrations between November 1993 and June 1995. From this total group, an analysis sample was chosen that met specified criteria. The purposes of the analysis sample criteria were to select examinees for whom the test is most intended and to select examinees who appeared to be trying on the test. The criteria were as follows:

- U.S. citizen
- English as best language
- reason for taking the GRE either a) application for graduate school, b) application for fellowship, or c) graduate department requirement
- no restarts in their testing session
- normally timed session
- no cancelled scores
- spent at least one-third of total testing time
- answered at least one-half of the questions in the section

Results from Actual CAT Data

Analyses were conducted to examine how examinees were proceeding through the CAT. First, information about the number of items answered by examinees was summarized by subgroup (Table 1) and by GRE score level (Table 2). In Table 1, for each measure most examinees answered all items, and the vast majority of examinees received scores. There are, however, more NSs in the analytical measure than in quantitative and verbal. Note that because of the relatively large NS rate the percent of examinees answering 27 items is listed for the analytical measure. The percent answering one less item than the minimum would indicate the extent to which examinees were almost able to get a score. The proportion answering 27 items was between 1-2 percent for all subgroups, indicating that most examinees who did not get an analytical score were more than one item short of the minimum number of items.

Overall, the subgroups show similar patterns, with the possible exceptions that a) the proportion of analytical NSs increases as age group increases, and b) Asian American test takers answer fewer quantitative and analytical items than other subgroups. This later finding may be a result of more difficult items being delivered to Asians because they tend to be of higher ability and get higher scores.

Table 1
Percent of Examinees Answering Various Numbers of CAT Items by Subgroup

	#of Items	T	F	M	AA	As	H	NA	W	<24	24-30	31-40	>40
V	<24	1	1	1	2	1	1	2	1	0+	1	1	2
	24	1	1	1	1	1	1	2	1	1	1	1	1
	25	2	2	2	1	2	2	1	2	2	2	2	2
	26	2	2	2	2	2	2	2	2	2	2	2	3
	27	2	2	2	2	2	2	3	2	2	2	2	2
	28	2	2	2	2	3	2	3	2	2	3	3	3
	29	4	4	4	5	4	4	2	4	4	4	4	4
	30	86	87	86	84	87	85	86	87	88	86	86	83
	N	70378	40435	29700	4973	2112	3018	590	57787	25241	24076	13431	7630
Q	<23	1	1	1	1	2	1	1	1	1	1	1	2
	23	2	2	3	1	3	2	2	2	2	2	2	2
	24	2	2	3	2	3	2	2	3	3	3	2	2
	25	3	2	3	2	4	3	2	3	3	3	3	3
	26	4	3	4	3	5	3	3	4	3	4	3	3
	27	6	5	7	4	7	5	6	6	5	6	6	5
	28	82	84	79	87	76	84	84	82	83	81	82	82
	N	70041	40220	29580	4920	2109	2998	586	57543	25179	23951	13337	7574
A	<28	4	4	4	4	5	4	4	4	3	4	5	8
	27	1	1	1	1	2	1	2	1	1	1	1	1
	28	5	4	5	3	5	4	3	5	5	5	4	4
	29	5	4	5	4	5	4	4	5	5	5	5	5
	30	7	7	7	6	6	6	7	7	7	7	7	7
	31	4	4	4	3	5	4	4	4	4	4	4	4
	32	4	4	4	3	4	4	3	4	4	4	5	4
	33	4	4	4	4	5	4	5	4	4	4	4	5
	34	7	7	8	7	9	7	7	7	7	8	8	7
	35	60	61	59	66	56	63	63	60	63	60	58	55
	N	67120	38529	28360	4762	2030	2851	561	55094	24287	22980	12657	7196

T=total analysis sample; F=female; M=male; AA=African American; As=Asian; H=Hispanic; NA=Native American; W=White; <24=less than 24 years of age; 24-30=24-30 years of age; 31-40=31-40 years of age; >40=greater than 40 years of age

Table 2
Percent of Examinees Answering Various Numbers of CAT Items by GRE Score Level

CAT	# of Items	200-290	300-390	400-490	500-590	600-690	700-800
V	24	1	1	1	1	1	1
	25	0+	2	2	2	2	1
	26	2	2	2	2	2	1
	27	1	2	2	2	2	1
	28	2	2	3	3	3	1
	29	3	4	4	4	4	2
	30	92	88	86	86	83	92
	N	542	11889	22095	21476	10651	3193
Q	23	0+	0+	1	2	3	7
	24	0+	1	1	2	4	6
	25	1	1	1	2	4	6
	26	0+	1	2	3	5	6
	27	1	2	4	5	9	10
	28	98	96	91	85	75	66
	N	1708	8764	15714	18217	15199	9631
A	28	1	2	2	4	6	8
	29	2	2	3	5	6	8
	30	2	3	5	7	9	11
	31	1	2	3	4	5	7
	32	1	3	3	4	5	6
	33	2	3	4	5	5	5
	34	4	5	7	8	9	8
	35	88	81	73	64	56	48
	N	1503	5921	11318	16467	16553	12659

In Table 2, for quantitative and analytical, the higher the GRE CAT score level, the smaller the proportion of examinees who completed all items. Also for these two measures, larger proportions of examinees at the three highest score levels (500 and above) answered exactly the minimum number of items to get a score. Among the possible explanations for this finding are a) because of their high ability level, these examinees are administered mostly difficult items that take more time to answer, and b) these examinees are intentionally not finishing and are using the test-taking strategy to maximize the average time allotted per item.

In addition to the number of items answered, an examination of distributions of times spent on the CAT section and on items may shed light on how examinees proceeded through the CATs. This analysis is tempered by the fact examinees are administered different items of varying difficulties and order, and therefore inferences about item times and even section times are tenuous. Nevertheless, some observations may be useful.

Tables 3 and 4 present summary statistics of times spent per CAT section. The summary statistics include mean, standard deviation and percent of the total allotted time as well as the number of examinees. Table 3 shows the CAT section times by demographic subgroup, and Table 4 shows the CAT section times by score level. These data show that a larger proportion of CAT section time was spent on the analytical measure than on the other two measures. In Table 3, the average time spent per section was similar across subgroups. In Table 4, the higher the ability level, generally the more time was spent on the section. One explanation of this is the higher ability levels are administered more difficult items, and more difficult items take longer to answer. Another possible explanation is that high ability examinees budget time well and use the strategy effectively.

Table 3

Summary of CAT Section Test Times in Minutes by Subgroup:
Mean, Standard Deviation, Mean as a Percent of Total Section Time, and Number of Examinees

CAT		T	F	M	AA	As	H	NA	W	<24	24-30	31-40	>40
V	Mean	26.4	26.2	26.7	26.4	26.6	26.6	26.1	26.4	26.1	26.5	26.5	26.6
	SD	3.6	3.6	3.5	3.8	3.4	3.5	3.9	3.5	3.7	3.5	3.5	3.5
	%	88	87	89	88	89	89	87	88	87	88	88	89
	N	70378	40435	29700	4973	2112	3018	590	57787	25241	24076	13431	7630
Q	Mean	38.6	37.8	39.9	36.8	40.4	38.5	37.1	38.8	38.7	39.0	38.4	37.9
	SD	6.9	7.2	6.4	7.7	6.1	7.0	7.5	6.9	6.8	6.9	7.2	7.2
	%	86	84	89	82	90	86	82	86	86	87	85	84
	N	70041	40220	29580	4920	2109	2998	586	57543	25179	23951	13337	7574
A	Mean	56.8	56.3	57.5	55.5	57.8	56.4	56.0	56.9	56.8	57.1	56.8	56.5
	SD	6.2	6.5	5.6	7.7	5.1	6.7	7.0	6.0	6.0	6.0	6.5	6.8
	%	95	94	96	93	96	94	93	95	95	95	95	94
	N	67120	38529	28360	4762	2030	2851	561	55094	24287	22980	12657	7196

Total CAT section times are 30 minutes for verbal, 45 minutes for quantitative, and 60 minutes for analytical.

T=total analysis sample; F=female; M=male; AA=African American; As=Asian; H=Hispanic; NA=Native American; W=White; <24=less than 24 years of age; 24-30=24-30 years of age; 31-40=31-40 years of age; >40=greater than 40 years of age

Table 4
Mean, Standard Deviation, and Percent of CAT Test Times in Minutes by GRE Score Level

CAT		200-290	300-390	400-490	500-590	600-690	700-800
V	M	24.0	25.9	26.6	26.7	26.4	25.3
	SD	5.1	3.9	3.4	3.4	3.5	3.9
	%	80	86	89	89	88	84
	N	542	11889	22095	21476	10651	3193
Q	M	29.8	33	36.4	39.1	41.6	42.9
	SD	7.9	7.8	7.0	6.1	4.7	3.6
	%	66	73	81	87	92	95
	N	1708	8764	15714	18217	15199	9631
A	M	46.3	52.0	55.4	57.1	58.3	58.7
	SD	12.9	9.5	7.0	5.3	3.8	3.0
	%	77	87	92	95	97	98
	N	1503	5921	11318	16467	16553	12659

Figures 1a - 1c present mean item times by item location in the CAT. For each measure, means are presented for examinees who answered all items and for examinees who answered exactly 80% of the items. These data are intended to assist in inferring how examinees proceed through the CAT. The spikes in the plots representing the largest mean item times are for the first items administered in sets, where the time spent reading the stimulus is confounded with the time spent on the first item in the set. Although there is no specified order of the delivery of item types, there are common patterns. Item times for the first item beyond the 80% threshold are provided because examinees could answer the 80% item, and then let time expire while they are on the next item without answering it. If examinees are using the strategy of answering 80% of the items and then stopping, then they would spent a relatively large amount of time on the item at the 80% threshold or the item one beyond. These data also shed light on the hypothesis that because of time constraints many examinees may be rushing to complete 80% of the items.

The data for examinees who answered exactly 80% of the items suggest that examinees generally are not using the strategy of answering 80% of the items and then letting time expire. The two plots on each figure are fairly parallel. This suggests that the 80% sample is working more slowly than the 100% sample throughout the test, and therefore the 80% sample does not seem to be reaching the 80% point with substantial time remaining and letting time expire. This also is illustrated at the 80% item and the first item beyond that. With one minor exception in the quantitative measure, the shortest average item time was always on the item at the 80% threshold. This suggests that on average these examinees work slowly and then need to rush at the very end to answer enough items to get a score.

For examinees who completed the verbal and quantitative sections, the mean item times were fairly similar throughout the section. For the analytical section, however, the three shortest mean item times were in item sequence numbers 32-34. This may indicate that many examinees were rushing to complete the analytical section, even though they did not need to complete the test to get a score.

Figure 1a

Verbal Mean Item Times (in seconds) by Item Position

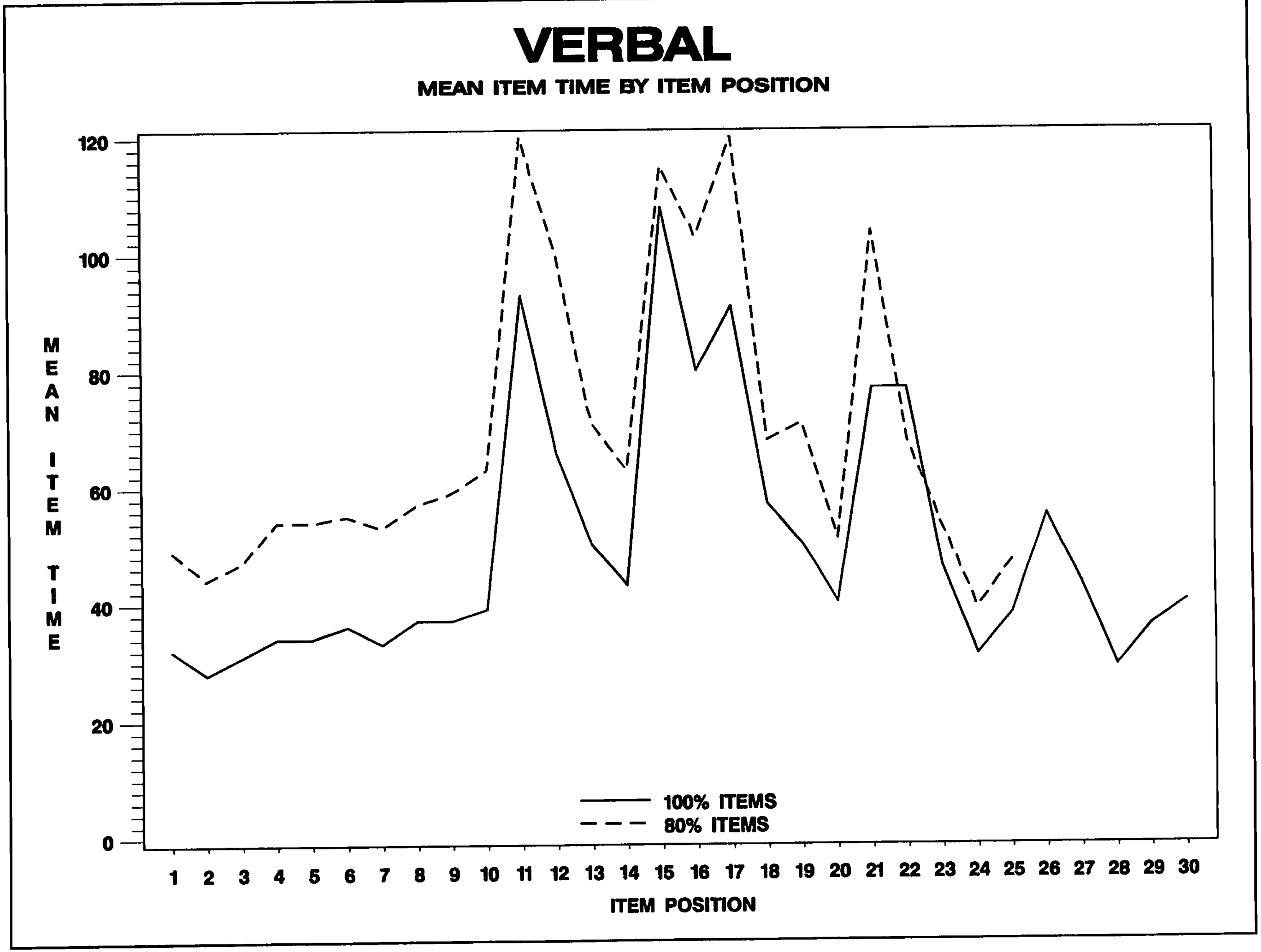


Figure 1b
Quantitative Mean Item Times (in seconds) by Item Position

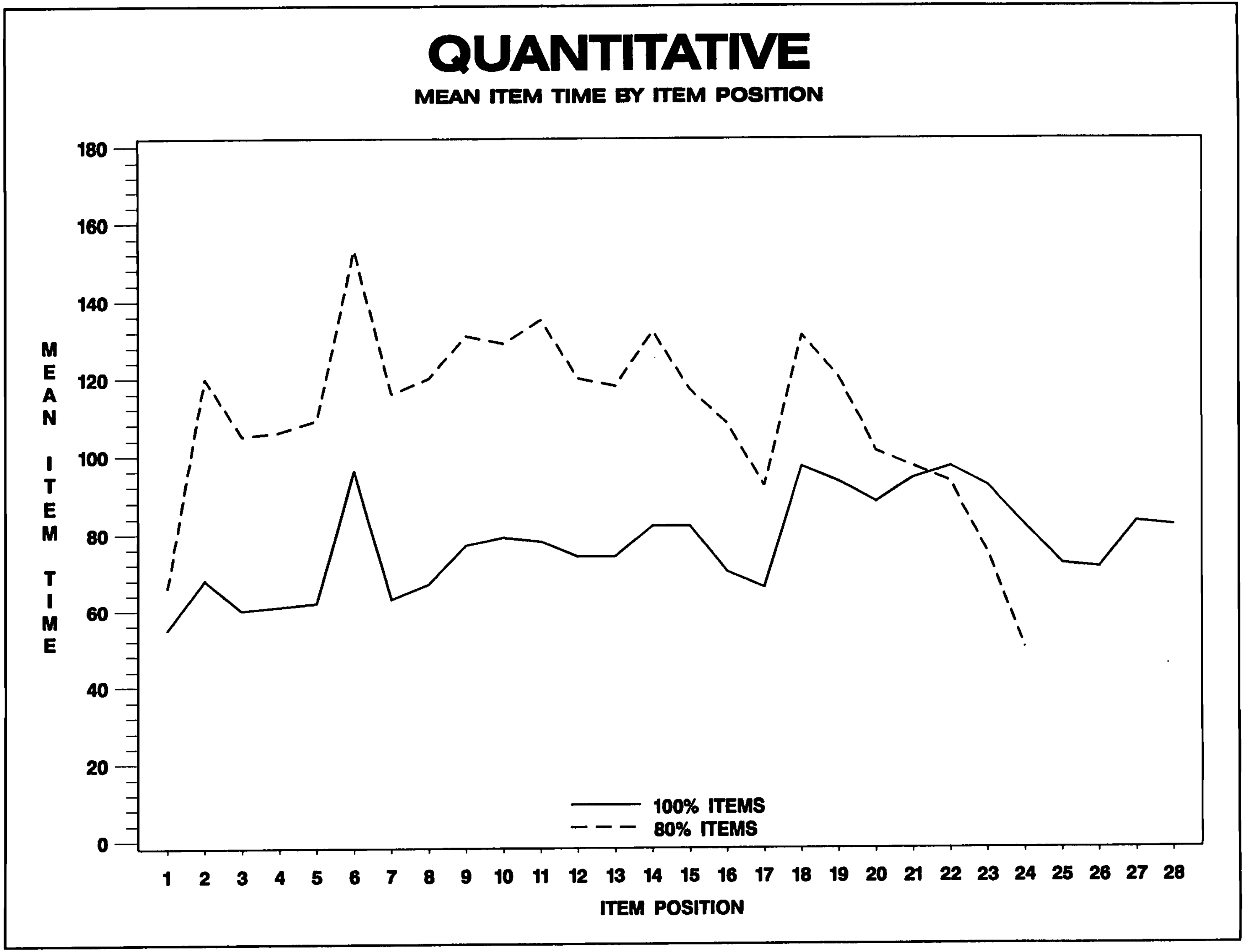
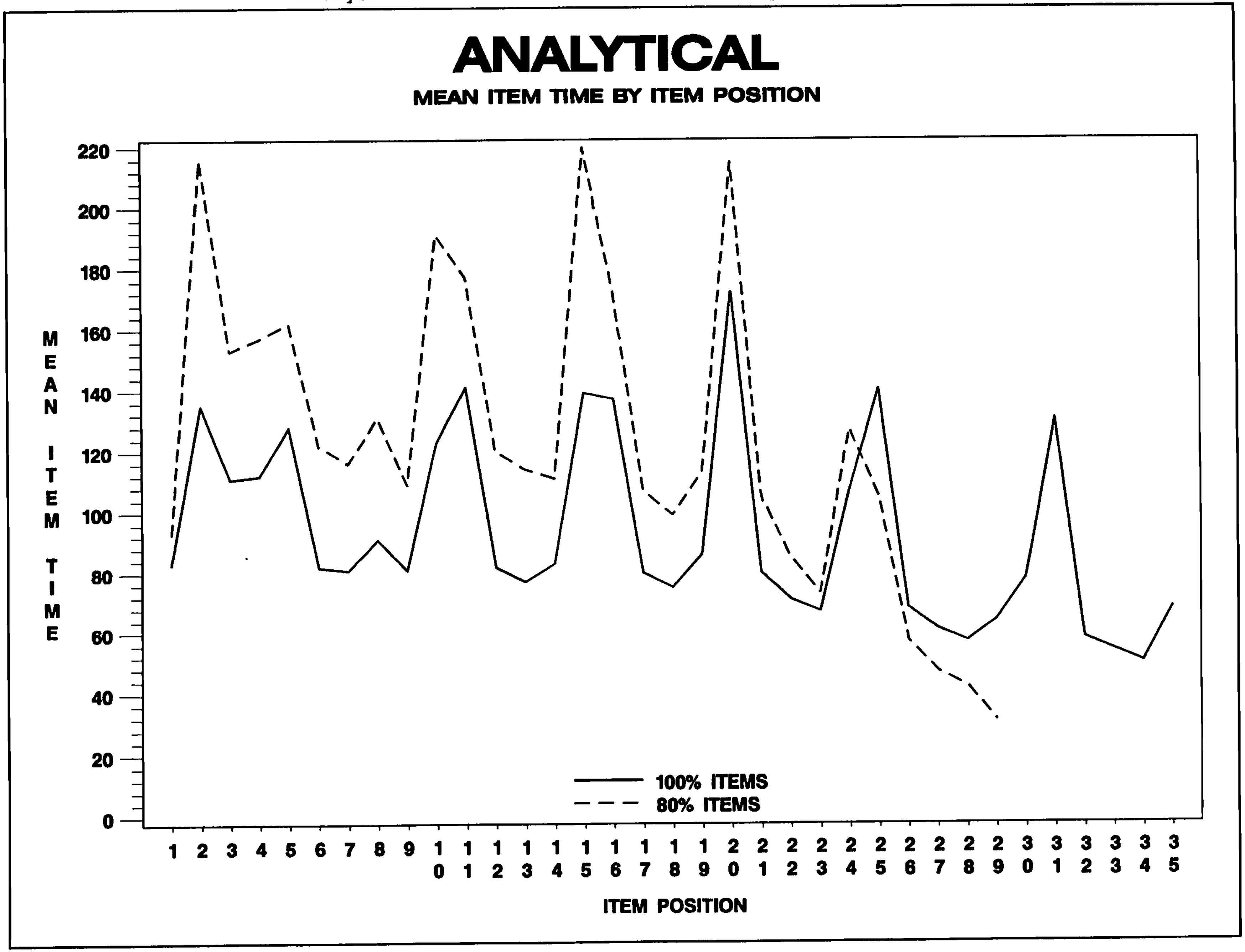


Figure 1c
Analytical Mean Item Times (in seconds) by Item Position



100% Minus 80% for Complete CATs

A sample of examinees was used to analyze the potential impact of using the strategy of stopping at the 80% point. That is, this analysis begins to address the question: What if examinees who completed the CAT had actually stopped at the 80% point? Examinees in this sample were administered one of two pools in April and June 1995 and had answered all items in their CAT. Their interim ability score at the 80% point was computed, and the difference in the final score and the score at the 80% point was evaluated. These data are presented in Table 5.

The largest differences were found for analytical, where the mean differences were about 7-19 points higher at the 80% point than at the end of the test. This trend was evident at all ability levels. Possible explanations include a) fatigue toward the end of the test lowered their scores, and b) examinees were rushing to finish the test and their haste lowered their scores. In addition, the potential impact may have been even greater if examinees had employed the strategy of using all of the allotted section time to answer only 80% of the items because they would have spent more time per item on average and thus may have performed better.

Table 5
Differences Between Scores at 80% and 100% (80% - 100%)
for Actual Data from Two Pools

		Verbal		Quantitative		Analytical	
Score Level		Pool 6	Pool 8	Pool 6	Pool 8	Pool 6	Pool 8
700-800	Mean	0-	4	1	2	7	4
	SD	14	16	15	14	19	21
	N	73	94	228	269	229	170
600-690	Mean	0+	6	0+	-2	14	14
	SD	15	16	21	19	27	33
	N	307	304	492	576	442	390
500-590	Mean	1	2	1	0-	14	10
	SD	18	18	21	22	30	32
	N	823	823	778	712	526	631
400-490	Mean	4	3	-1	1	19	12
	SD	20	21	22	22	38	32
	N	1072	1085	914	837	616	616
300-390	Mean	4	3	0+	1	14	17
	SD	15	16	20	16	36	39
	N	737	778	612	670	337	393
200-290	Mean	-2	1	1	0-	11	9
	SD	12	16	19	18	41	35
	N	59	60	154	151	161	180

Methods for Simulations

CAT simulations were conducted to assess the potential impact of different test-taking strategies and of different ways of scoring tests when not all items were answered. Interim scores as simulees proceeded through the CAT under various conditions were analyzed. Interim scores were obtained by converting the item-level ability estimate to the GRE scale. A total of 11 operational CAT pools were analyzed to provide an indication of the variability in results across different pools. Results were obtained for a total of 1000 simulees at each specified true ability level. The ability levels used in these analyses were the same as those used in the operational simulations. The numbers of ability levels was 13 for verbal, 11 for quantitative, and 9 for analytical.

First, scores at the end of the CAT were compared with scores that would result at the 80% point. This provides information on any bias that may be present in the CAT algorithm that may have an impact on the strategy of stopping at the 80% point. Second, scores at the 80% point were compared with scores at the item immediately preceding the first item marked wrong after the 80% point. This analyses illustrates the benefit of continuing beyond the 80% point until the examinee reaches an item they cannot answer correctly. Note that a weakness in these analyses is that the simulations cannot take into account the relationship between time spent per item and the probability of answering the item correctly. That is, the algorithm and thus the interim score does not account for the additional time per item that examinees could spend if they answer fewer than the total number of items.

CAT simulations also were conducted to assess the impact of implementing two different scoring rules for incomplete CATs. One rule is to mark all unanswered items beyond the 80% point as wrong. In the other rule, the probability of correctly answering each item after the 80% point is set to be random. These rules would serve to encourage examinees to complete the test and would provide a score penalty for those not finishing the test.

Results from Simulations

100% Minus 80% for Complete CATs. Simulations for complete CATs were used to assess the extent of any bias in the CAT algorithm in estimating ability at the 80% point versus at the end of the test. Mean differences between the final GRE score and the interim GRE score at the 80% point were computed for the various ability levels. The results indicated that the mean bias was negligible (i.e., the absolute value of the mean bias was less than 3 scaled score points) at virtually all ability levels and pools for each of the measures. However, The spread of differences indicated that for a very small percentage of simulees the scores could vary by 100 or more points in either direction between the 80% and the end of the test.

80+% Minus 80%. Simulation data were used to address the impact of the potential test-taking strategy of answering 80% of the items and then continuing only until an item is administered to which the examinee does not know the correct response. Because the only characteristic of the simulee-item interaction is the right or wrong response, for this analysis the score at the 80% point was compared with the score at the last consecutive item right after the 80% point. That is, this analyses assumes that examinees know whether they will get an item right or wrong, and will not answer any items once they are administered an item that they know they will get wrong. In a practical sense, this means that examinees, once they are administered an item beyond the 80% point that they know they will get wrong, do not answer it (or any more items) and let time expire (so they can get a score). This last consecutive item right beyond the 80% point will be labelled the 80+% point.

Distributions of the consecutive number of questions answered correctly beyond the 80% point were computed for each specified ability level for each pool. The differences across pools were minimal, so results are presented for all pools combined. The higher ability levels tended to get longer strings of consecutive right answers after the 80% point. It also can be seen that a large proportion of low ability simulees got the next item after the 80% point wrong and therefore for these simulees this strategy collapses to the strategy of stopping at the 80% point.

Tables 6a - 6c show for each measure and ability level a) the mean number of items answered correctly beyond the 80% point, b) the percent of simulees who got at least one item correct beyond the 80% point, c) the mean scale score gain between the point beyond 80% (labelled 80+%) and the 80% point, and d) the maximum score gain. The patterns were similar for all three measures. The mean consecutive number correct beyond the 80% point increased with ability level, and correspondingly the percent answering the next item after the 80% point correctly also increased with ability level. The mean score gains (80+% - 80%) were largest in the middle of the ability distribution and were in the 7-9 point range. The largest maximum score gains using this strategy were around 140-160 points.

Table 6a

VERBAL Consecutive Items Correct Beyond the 80% Point for Simulees in All Pools Combined*

THETA	TRUE SCALED SCORE	Mean Number Correct Beyond 80%	Percent Simulees with ≥ 1 Item Correct Beyond 80%	Mean Score Gain (80+% - 80%)	Maximum Score Gain
-5.86	220	0.1	12	0.8	40
-3.36	270	0.4	28	2.0	60
-2.34	310	0.7	44	2.6	50
-1.64	350	1.1	56	4.3	50
-1.07	390	1.4	62	6.4	80
-0.59	450	1.6	62	7.6	140
-0.13	500	1.7	63	8.2	140
0.33	550	2.0	67	7.8	70
0.80	610	2.3	74	7.8	60
1.30	660	2.8	82	8.0	60
1.86	720	3.5	87	7.4	50
2.61	780	4.6	94	5.2	60
4.88	800	5.8	99	0.1	20

* There are 6 items beyond the 80% point.

Table 6b

QUANTITATIVE Consecutive Items Correct Beyond the 80% Point for Simulees in All Pools Combined*

THETA	TRUE SCALED SCORE	Mean Number Correct Beyond 80%	Percent Simulees with ≥ 1 Item Correct beyond 80%	Mean Score Gain (80+% - 80%)	Maximum Score Gain
-3.84	200	0.1	11	1.0	70
-2.18	290	0.4	28	4.3	120
-1.38	370	0.7	42	6.9	160
-0.81	440	1.0	51	8.7	140
-0.35	500	1.2	53	8.7	120
0.05	560	1.3	56	8.2	90
0.43	620	1.4	61	7.9	100
0.81	670	1.6	64	7.3	70
1.24	720	1.9	71	6.0	50
1.88	780	2.8	80	3.9	40
3.55	800	4.6	97	0.0+	20

* There are 5 items beyond the 80% point.

Table 6c

ANALYTICAL Consecutive Items Correct Beyond the 80% Point for Simulees in All Pools Combined*

THETA	TRUE SCALED SCORE	Mean Number Correct Beyond 80%	Percent Simulees with ≥ 1 Item Correct Beyond 80%	Mean Score Gain (80+% - 80%)	Maximum Score Gain
-3.98	220	0.3	22	0.6	40
-2.23	280	0.5	32	3.0	150
-1.46	360	0.7	38	5.2	100
-0.86	430	0.9	45	6.5	90
-0.31	510	1.2	53	7.1	90
0.23	570	1.5	60	7.4	80
0.82	640	1.7	65	7.5	70
1.57	730	2.1	71	7.1	70
3.09	800	4.9	89	0.9	30

* There are 7 items beyond the 80% point.

Mark Wrong Beyond 80%. Table 7 shows the results of the scoring rule where unanswered items beyond the 80% point are marked wrong. The impact of the scoring rule is represented as the difference between the score at the 80% point and the score at the end of the test where all items after the 80% point have been marked wrong. The final column of the tables shows the average difference for each ability level across all 11 pools. The mean differences varied across ability levels. For verbal, the largest mean differences were between around 80-100 points at the upper half of the score scale. For quantitative, the largest mean differences were around 50-55 points and occurred for scores around 400-700. For analytical, the largest mean differences were around 70 points and occurred for scores around 500-700.

For all three measures, the standard deviations were rather large. This is due to the nature of the item selection and scoring algorithms. For each simulee toward the end of the test, different items generally are selected to be marked wrong. Because the items have different parameters, different scores result. Thus, the magnitude of the penalty could be quite different for examinees at the same true score level.

Mark at Random Beyond 80%. Table 8 shows the results of the scoring rule where unanswered items beyond the 80% point are marked at random. The impact of this scoring rule is represented as the difference between the score at the 80% point and the score at the end of the test where all items after the 80% point have been at random. To accomplish this, the probability of a correct response was set at .20 for items in the verbal and analytical measures because they are all 5-choice items. For the quantitative measure, the probability of a correct response was set at .225 because half of the items are 4-choice and half are 5-choice. The final column of the tables shows the average difference for each ability level across the 11 pools. The mean differences for this rule were smaller than for the rule where all unanswered items were marked wrong because some of these items beyond the 80% point were marked correct. As with the rule marking unanswered items wrong, the mean score differences varied across score levels, and the standard deviations of the differences were rather large.

Table 7

Mean Scaled Score Differences Between 80% and 100% (80% - 100%)*
Probability of unanswered items correct = .00

VERBAL

TRUE SCORE	MEAN DIFF	SD DIFF
220	4	8
270	11	10
310	18	12
350	32	18
390	50	24
450	65	28
500	76	32
550	82	34
610	83	33
660	89	34
720	95	34
780	101	33
800	90	30

QUANTITATIVE

TRUE SCORE	MEAN DIFF	SD DIFF
200	6	12
290	23	17
370	38	23
440	51	26
500	55	26
560	55	25
620	54	23
670	52	19
720	48	16
780	39	12
800	14	9

ANALYTICAL

TRUE SCORE	MEAN DIFF	SD DIFF
220	4	9
280	24	20
360	47	26
430	62	29
510	73	32
570	75	35
640	69	31
730	69	28
800	42	36

* Mean differences are for 11 pools combined.

Table 8

Mean Scaled Score Differences Between 80% and 100% (80% - 100%)*
Probability of unanswered items correct = Random (.20 for Verbal and Analytical;
.225 for Quantitative)

VERBAL

TRUE SCORE	MEAN DIFF	SD DIFF
220	-4	12
270	2	11
310	10	12
350	21	17
390	34	22
450	44	27
500	52	31
550	56	31
610	58	31
660	64	32
720	71	34
780	79	33
800	70	31

QUANTITATIVE

TRUE SCORE	MEAN DIFF	SD DIFF
200	-6	17
290	7	20
370	19	24
440	28	26
500	32	26
560	33	24
620	33	23
670	33	21
720	31	18
780	27	14
800	7	9

ANALYTICAL

TRUE SCORE	MEAN DIFF	SD DIFF
220	-1	9
280	11	18
360	27	24
430	38	27
510	47	29
570	49	31
640	47	27
730	49	23
800	30	17

* Mean differences are for 11 pools combined.

Summary of Results and Conclusions

This investigation addressed a number of issues related to the 80% rule and potential test-taking strategies. Although the 80% rule was originally intended to accommodate test-takers who naturally proceed slowly through the test, the 80% rule may be used by other test-takers as part of a test-taking strategy where they would pace themselves and intentionally not answer the last items in the section. To address the extent to which these kinds of test-taking strategies are being employed, CAT data from over 72,000 examinees who tested between November 1993 and May 1995 were used. Simulation data were used to determine the potential impact of several test-taking strategies and to determine the impact of imposing penalties to circumvent strategies and encourage examinees to answer all items. Some of the findings can be summarized as follows:

1. The distributions of the numbers of items answered were similar for ethnic and gender subgroups (with a few minor exceptions). For this sample of examinees, the no-score rates were about 1% for the verbal and quantitative measures, and about 4% for the analytical measure. The proportions of examinees who answered all items were 86% for verbal, 82% for quantitative, and 60% for analytical.

2. The distributions of the numbers of items answered differed by score levels for the quantitative and analytical measures. For these two measures, examinees with higher scores answered fewer items on average. Higher ability examinees tend to be administered harder items, and the harder items may take longer to answer. For the verbal measure, there was no appreciable differences in numbers of items answered by score levels.

3. Analyses of mean times spent on items did not indicate that many examinees were using the strategy of answering 80% of the items and then letting time expire.

4. For examinees who answered all the analytical items, the shortest item times were spent on items toward the end of the section. This suggests that examinees may be rushing to finish the test, even though they have already answered enough items to get a score. There were no trends based on mean item times evident in the verbal and quantitative sections.

5. For examinees who answered all the items in the section, their final scores were compared to what their scores would have been if they had stopped at the 80% point. The mean score differences were negligible for the verbal and quantitative measures. However, for the analytical measure at all score levels except the extremes, examinee scores were about 15 points higher on average at the 80% point than at the end of the test. Examinees may be rushing to finish the section and are making more errors (see previous paragraph). This appears to illustrate the impact of speededness on scores. This also suggests that with all other factors held constant, such as the apparent speededness of the measure, on average examinees would receive higher analytical scores if they answer 80% of the items than if they answer all items.

6. Simulations were conducted to assess the potential impact of the test-taking strategy of answering 80% of the items and then continuing to answer items until

an item is administered to which the examinee doesn't know the answer. The patterns of results of employing this strategy were similar for the three measures. The mean score gains using this strategy versus stopping at the 80% point were largest in the middle of the ability distributions and were about 8 points.

7. To address the concern that examinees may intentionally answer only 80% of the items in a section, simulations were conducted to assess the impact on scores of employing two different scoring rules for incomplete CATs. With one scoring rule, unanswered items in the section are marked as wrong. With the other scoring rule, unanswered items are marked at random. Simulated distributions of mean score differences were produced between scores at the 80% point and at the end of the test after invoking the scoring rule. When unanswered items were marked wrong, the mean differences at the middle ability levels ranged from about 50-90 points for verbal, 40-55 points for quantitative, and 50-75 points for analytical. The standard deviations of the score differences were rather large (about 25-35 points) due to the nature of the item selection and scoring algorithms. The mean score differences for the rule where items were marked at random were somewhat smaller than those for the rule where items are marked wrong, but the standard deviations of the score differences for the two rules were about the same. The rather large standard deviations indicate that examinees with very similar true scores may end up with very different penalties being assessed as a result of one of these scoring rules.

Rules for scoring incomplete CATs may take many forms. This paper looked at the 80% rule and at finishing incomplete CATs with wrong or random responses to items selected using the updated ability estimate. Another possibility would be to finish an examinee's incomplete CAT by marking wrong or at random items selected based on the examinee's ability estimate at the time the examinee actually stopped answering items. Or, a penalty could be assessed that is equal to a fixed number of score points for every unanswered item. Another alternative would be to base the magnitude of the penalty on the reliability of the incomplete CAT.

To date, the 80% rule has worked reasonably well. These data do not indicate that there is currently a widespread use of the 80% rule as a strategy for the GRE CAT. The program will need to continue to monitor examinee performance since raise serious psychometric and equity issues would be raised if examinees begin to employ the strategy.

Establishing a rule for scoring incomplete CATs requires considerable care. Two types of examinees that may not finish the test include: a) examinees who implement a test-taking strategy of intentionally stopping before the end of the test in an attempt to maximize their score, and b) examinees who spend (and need) more time actually reasoning through the items. In an ideal psychometric world, all examinees would answer every item. But given that not all examinees finish the test, an ideal scoring rule would penalize the first type of examinee and not the second. Unfortunately, it may not be possible to differentiate these two groups of examinees with a high level of accuracy. There is no perfect solution to scoring incomplete CATs. Objectives must be weighed and tradeoffs must be considered when deciding upon a scoring rule. Ultimately the selection of a scoring rule for incomplete CATs is a matter of policy. Any scoring rule, however, should be revisited periodically and modified if necessary in light of examinee behavior.

NCME Annual Meeting, April 9-11, 1996

TM025880

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Computing Scores for Incomplete GRE General Computer Adaptive Tests

Author(s): Sharon C. Slater & Gary A. Schaeffer

Corporate Source: Educational Testing Service

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ **Sample sticker to be affixed to document**

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

_____ Sample _____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document □

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

_____ Sample _____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Sharon C. Slater	Position: graduate student
Printed Name: Sharon C. Slater	Organization: Univ of Massachusetts, Amherst
Address: 159 Hills South / REMP University of Massachusetts Amherst, MA 01003	Telephone Number: (413) 545-1947 Date: 5/22/96

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

March 12, 1996

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (23)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

NCME 1996/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the NCME web page (http://www.assessment.iupui.edu/ncme/ncme.html). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an NCME chair or discussant, please save this form for future use.