Item response latencies of different item formats for ethnic groups matched on ability.

Sharon C. Slater

University of Massachusetts Amherst
ITEM RESPONSE LATENCIES OF DIFFERENT ITEM FORMATS
FOR ETHNIC GROUPS MATCHED ON ABILITY

A Thesis Presented
by
SHARON C. SLATER

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of
MASTER OF SCIENCE
February 1996
Psychology
ITEM RESPONSE LATENCIES OF DIFFERENT ITEM FORMATS FOR ETHNIC GROUPS MATCHED ON ABILITY

A Thesis Presented
by
SHARON C. SLATER

Approved as to style and content by:

James M. Royer, Co-chair

Ronald K. Hambleton, Co-chair

Hariharan Swaminathan, Member

Arnold Well, Member

Melinda A. Novak, Department Head Psychology
ACKNOWLEDGEMENTS

I would like to acknowledge both the Educational Testing Service (ETS) and the Graduate Record Examination (GRE) Program for a number of opportunities they have provided over the last two years. The GRE Assistantship in Psychometrics program provided funding and valuable experiences, and exposed me to many issues in large scale assessment and computerized testing. I am also grateful to ETS and the GRE Program for providing the data used in this study. Craig Mills and Maria Potenza were especially helpful and supportive.

I would like to thank the co-chairs of my thesis committee, Professors Ronald Hambleton and Mike Royer, for their continuous support and technical guidance. Professors H. Swaminathan and Arnie Well were especially helpful with the data analyses; and I am appreciative of their suggestions and advice.
ABSTRACT

ITEM RESPONSE LATENCIES OF DIFFERENT ITEM FORMATS
FOR ETHNIC GROUPS MATCHED ON ABILITY

FEBRUARY 1996

SHARON C. SLATER, B.S., JUNIATA COLLEGE
M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professors James M. Royer, and Ronald K. Hambleton

Item response latencies for eight item types from the Graduate Record Examination (GRE) were compared for black and white examinees who had been matched on ability. Differences in mean item response latencies between the two groups were found for only two of the eight item types. One of the differences could be explained by a confound introduced by the matching procedure; the other difference is harder to explain. Also, some interesting patterns of item response latencies were identified for examinees of different ability levels and for correct and incorrect responses. Future research could attempt to further explain why differences in item response latencies were found by interviewing students, either during or after testing. Overall, the findings suggest little evidence of differences in mean item response latencies between black and white groups, when the groups are matched on ability.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
</tbody>
</table>

**Chapter**

1. REVIEW OF THE LITERATURE .......................... 1  
   Test Level Response Time Research ................. 2  
   Individual Differences ............................ 2  
   Group Differences ................................ 6  
   Item Level Response Time Research ................. 10  
   Summary ........................................... 23  

2. PURPOSE OF THE STUDY ............................... 26  

3. RESEARCH METHOD .................................. 29  
   Examinees ....................................... 29  
   Graduate Record Examination Computer Based Test .. 30  
   Verbal Section .................................. 31  
   Quantitative Section ............................. 34  
   Analytical Section ............................... 36  
   Procedure ....................................... 37  

4. RESULTS .......................................... 44  
   Main Effect Due to Ethnicity ....................... 44  
   Main Effect Due to Ability ....................... 45  
   Regression Analyses ............................. 48
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Means and Standard Deviations of GRE-CBT Subtest Scores for Matched Groups</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of Examinees with Effect Sizes of the Differences Between Groups</td>
<td>39</td>
</tr>
<tr>
<td>2.</td>
<td>Means and Standard Deviations of GRE-CBT Scores</td>
<td>40</td>
</tr>
<tr>
<td>3.</td>
<td>Mean Item Response Times and Standard Deviations for Different Item Types</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>for Black and White Examinees</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Mean Item Response Times for Different Item Types by Ability Level</td>
<td>46</td>
</tr>
<tr>
<td>5.</td>
<td>Mean Item Response Times for Different Item Types by Correctness of Response</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>for Black and White Examinees</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Comparison of Mean Item Response Times and P-Values for Analytical and</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Logical Reasoning Items in the First and Second Halves of the Analytical</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Section</td>
<td></td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Response Times of Black and White Examinees Across GRE Verbal Score</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>for Sentence Completion Items</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Response Times of Black and White Examinees Across GRE Analytical Score</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>for Logical Reasoning Items</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Response Times of Black and White Examinees Across GRE Analytical Score</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>for Analytical Reasoning Items</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Response Times of Black and White Examinees Across GRE Verbal Score</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>for Reading Comprehension Items</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Response Times of Black and White Examinees Across GRE Quantitative Score</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>for Quantitative Comparison Items</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Response Times of Black and White Examinees Across GRE Quantitative Score</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>for Problem Solving Items</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1

REVIEW OF THE LITERATURE

Speededness is an ongoing concern in testing. If a test is found to be speeded, the validity may be called into question. If speeded, the test is no longer just a measure of the examinee's command of the content, but also of how quickly the examinee can respond to questions about the content. Speededness also has implications for the fairness of a test. If one subgroup of the population needs significantly more time to complete the test than another, the test scores may be biased against the subgroup that did not have sufficient time to finish the exam.

A number of approaches for investigating speededness at the test level have been applied to various tests (Evans & Reilly, 1972; Lord, 1956; Wild, Durso & Rubin, 1982). These involve giving examinees more time to complete the entire test and looking for differences in scores for the original test and the test with extended time limits.

Speededness has also been examined at the item level. Recently, with the expanded use of computers in testing, it has become possible to obtain much more accurate
measures of how long an examinee takes to respond to each individual item in a test. These "response latencies" can be used to investigate issues of speededness or bias at the item level, as well as a number of other issues.

The literature review will first describe the study of speededness at the test level. The research conducted in this area for investigating both individual and group differences will be described, and shown to be inconclusive. Next, a number of reasons why the study of timing at the item level is important will be discussed. Finally, the current research on response latencies will be summarized and the purpose of the present study will be introduced.

Test Level Response Time Research

Research on response times at the test level can be divided into two parts. In the first section, the research focused on differences between individuals will be described. The second section will incorporate studies that compared groups to identify differences.

Individual Differences

Reaction time has been a variable of interest with regards to individual differences for over 100 years (Jensen, 1982). In 1862, Francis Galton was the first to
formally suggest that individual differences would be reflected in reaction times. He postulated that the more mental ability one has, the quicker one's response times would be. In his research, he did not find the differences he had expected and abandoned the effort. However, a student working with Galton, James Cattell, did continue work in the area of individual differences in reaction times and brought the idea to America. Since that time, many studies on reaction time and "intelligence" have been conducted (Baxter, 1941; Jensen, 1982; Lord, 1956; Spearman, 1927; Thurstone, 1937).

Response time has been of particular interest in testing for some time. In 1937, Thurstone presented a theoretical model to attempt to separate the effects of ability, motivation and speed in mental tasks. This model used what he called a "psychometric surface" (a three dimensional space defined by response time, probability of a correct response, and difficulty) to illustrate the relationships of the variables. The model described claimed to be able to appraise mental ability independently of speed of performance. Motivation was also found to be independent of ability. Thurstone (1937) stated that as long as the level of motivation has a
positive value, motivation will not affect the person's probability of a correct response. It will only enable the person to respond to the item more quickly. Zero or negative motivation would translate into indifference or aversion to the task, respectively, and would result in no response at all. Thurstone's (1937) model showed that mental ability, independent of speed and motivation, could be measured experimentally. However, the model has little practical use because mental ability is defined as "that degree of difficulty for which the probability is 1/2 that he will complete the task in infinite time" (Thurstone, 1937). Rarely, if ever, does an examinee have infinite time to complete an item. Thurstone does present an explanation of how mental ability can be interpolated by choosing a number of points on the psychometric surface for different values of response time and different values of difficulty. Curves parallel to the plane of the probability of success and response time are determined for several difficulty values. The difficulty value whose cumulative frequency curve has an asymptotic limit of the probability of success equal to 0.5 is considered to be the examinee's mental ability. This complicated procedure would have to be conducted on individual examinees and
would be too difficult and time consuming to be of much practical use.

In another investigation of speed and ability, Lord (1956) investigated speed factors in tests and academic grades of students at Annapolis. Through factor analysis, four speed factors were identified: a number-speed factor, a perceptual-speed factor, a verbal-speed factor, and a spatial-speed factor. Five course grades were included in the analyses: English, Foreign Language, Engineering Drawing and Descriptive Geometry, Chemistry, and Mathematics. All correlations between course grades and the four speed factors were positive, except the correlation between English and the spatial-speed factor. Although the correlations were not large, Lord concluded that various kinds of speed play at least some part in the course grades studied. For this reason, he stated that "speededness in admissions tests is to this extent justified."

Despite the many studies conducted on response time and ability, the results have been inconclusive. Rindler (1979) cites many studies that suggest there is "neither a strong nor consistent relationship between response time and response accuracy," (i.e., ability). Himmilweit
(1946) was even able to show that rate of response was more associated with examinee temperament than with the ability to respond correctly. It seems clear that researchers have not been able to determine exactly what, if anything, the study of response times can tell us. However, research continues (Bergstrom, Gershon & Lunz, 1994; Kingsbury, Zara & Houser, 1993, 1994; Llabre & Froman, 1987; Van de Vijver & Helms-Lorenz, 1994).

Group Differences

Comparisons among different subgroups of the population are often made in many lines of research. The study of response times is no different. The research investigating response times for individual differences inevitably led to research investigating response times for group differences (e.g., male/female, black/white). Daly and Stahmann (1968) studied the effect of time limits on a university placement test for English. The subjects in this study were college freshmen who had been assigned to a remedial English class based on their scores on the Cooperative English Expression Test. Before the start of the class, the test was readministered to the students with extended time limits. The test was then administered a third time at the end of the remedial course.
Daly & Stahmann (1968) found that additional testing time, essentially, had the same effect on the effective usage of English as the remedial English course. In other words, the students scores improved significantly when test time limits were increased; the remedial course that immediately followed did not significantly improve the scores in effective usage of English beyond the improvement achieved with additional time. (The course focused more on the mechanics (e.g., grammar, punctuation) of English, and the scores in this area did improve significantly beyond the effect of additional time.) Daly and Stahmann (1968) also reported that had the scores from the extended time-limit test been used for placement, 41% of the sample would not have been placed in remedial English.

In another study of speededness and ethnicity, Durbin, Osburn and Winick (1969) compared test response time (i.e., completion of the test) of blacks and whites to determine if testing procedures were serving to discriminate against minority examinees. They focused on testing time to determine if highly speeded tests were equally fair to blacks and whites. High school students of low and high socio-economic status from both racial
groups were tested. The results indicated that all groups improved their scores in a comparable way; therefore, Durbin et al. concluded that the testing procedure did not discriminate between racial groups or between culturally advantaged or disadvantaged groups.

Along the same lines, Evans and Reilly (1972) investigated speededness as a source of test bias for fee-free and regular candidates on the Law School Admissions Test (LSAT). Fee-free candidates are those for which the cost of testing is waived. In this study the fee-free candidates were selected from a sample of predominantly black colleges. Students from both the fee-free and regular candidate groups were administered either the usual "speeded" form, or a special form with the same time limits, but fewer items. Results found that, based on Swineford's 1956 criteria for speededness (cited in Evans & Reilly, 1972) used at ETS, "the Reading Comprehension sections of the LSAT would be considered a speeded measure for fee-free candidates and an unspeeded measure for regular center candidates."

However, a nonsignificant interaction between group (fee-free vs. regular) and test (speeded vs. unspeeded) indicated that reducing the speededness was not
significantly more beneficial (in terms of increasing the number of items answered correctly) to fee-free than to regular center candidates. Therefore, they concluded that the time limits on the LSAT did not serve to discriminate against fee-free candidates.

Following Evans and Reilly (1972) in the investigation of speededness for different ethnic groups on large-scale standardized assessments, Wild, Durso, and Rubin (1982) investigated time limits on the Graduate Record Examination (GRE). Ten minutes were added to both the quantitative and verbal experimental sections. The effects of the additional time were studied in a number of comparisons: black/white, male/female, and groups differing in the number of years since they had earned their baccalaureate degree. The results from this study also found that additional time did not differentially improve scores for any of the groups studied.

The results of the studies described above are inconclusive. Some found that test time limits do discriminate against certain groups; some found that the time restraints do not discriminate against certain groups. Others found that although the tests were indeed speeded for some groups, the test time limits did not
discriminate against any group studied, because increasing the time limits for all examinees did not differentially improve the scores of any group. It is clear that more research on speededness is still needed.

**Item Level Response Time Research**

All of the studies described in the previous section dealt with response time at the test level, i.e., the amount of time allotted for the completion of the entire test. In these studies, there was no way of knowing how examinees allocated their time throughout the test. There is no way to determine timing behavior for individual items. More precise item timing information could provide more accurate information for speededness research.

The investigation of response times at the individual item level has previously been difficult, if not impossible. The technology to easily collect accurate response time data was simply not available. Some researchers tried to assess item response time by observing each examinee and timing the responses to each item with a stopwatch (Tate, 1948). This method is not only inefficient, but has limited generalizability to real testing situations.
Another approach is to divide the total test time by the total number of items on a test to determine the amount of time expected to be spent on each item (Wild, Durso, & Rubin, 1982). This method is also unsatisfactory, because all items on a test rarely, if ever, take an equal amount of time to solve. This inability to obtain timing information at the item level placed an obvious limitation on the kinds of analyses which could be conducted on individual response latencies.

Recently, due to the expanded use of computerized testing, response latencies at the item level are more readily available. Response latency generally is defined as the total time that it takes for a person to respond to a test question once it is presented on the computer screen (Kingsbury, Zara, & Houser, 1993). With this information, researchers have a new tool that enables the investigation of a number of issues that could not be addressed at the test level. Many of these issues will be discussed below.

As mentioned, the availability of item response times is relatively new. Therefore, the first issue in studying response latencies is to gain insights into the ways examinees approach test taking situations. By looking at
patterns of response times for examinees, information on how test-takers allocate their time on a test can be gained. Also, by looking at a number of examinee and item variables, research can be done to determine which variables influence response times.

Beyond this type of exploratory research, there are a number of additional reasons why researchers are interested in studying response times. For one, there are important uses of item timing information in the area of test construction. With specific response time information for items, it would be possible to construct a test using more or less time consuming items, depending on the needs of the test. For example, if a low discriminating item is found to be very time consuming, then it might be wise to exclude it from the test, if content specifications allow.

Secondly, response latencies could aid in developing test administration policies. Information about item response times could provide a clearer prediction of how much time is needed for examinees to complete a test constructed of certain items. This would enable realistic time limits to be set for the test. Time limits on achievement tests, in theory, should not be an issue.
because they are generally considered to be power tests. However, practicality usually dictates that time limits must be set, especially for expensive computer time needed for tests administered by computer. Kingsbury et al. (1994) caution against this use of response latencies in computer adaptive testing, stating that it "perpetuates bad testing practices."

Third, item level response time information could be used to identify aberrant response patterns that might indicate an examinee with low motivation or one who is cheating. It is often difficult to identify examinees who exhibit these patterns in adaptive tests, because the test tends to correct for unusual responses on an item-by-item basis. If an examinee is responding to a series of items in far less time than most examinees spend on the item, this may be an indication of a problem.

A fourth reason way in which response latencies could be used would be to gain information about the life of items. More specifically, Kingsbury et al. (1994) see response latencies as potentially able to provide clues to item overexposure. Overexposed items would have a pattern of quicker response rates across many different examinees. These items would function similarly to items
cheated on by an examinee, but the effect would be more wide-spread across a greater number of examinees.

Fifth, response latencies could be used in an educational setting to identify areas where a student may be having difficulties. If a student responds more slowly to a certain type of item, it may indicate a problem with that particular item format. Or, more importantly, if a student consistently responds more slowly to items of a certain content, it may indicate insufficient knowledge of that area. This is opposite to the findings of Kingsbury et al. (1993) and Schaeffer et al. (1993) who found that low ability examinees actually have quicker responses to test items of greater difficulty.

However, the testing situation differs from an educational context. In a test, an examinee cannot afford to spend too much time on any one item, and is often instructed to guess because there is no penalty for incorrect responses. In an educational setting, the student is not under strict time constraints and the emphasis is on learning the material, not necessarily on quickly determining the correct response.

A sixth reason for studying item response latencies is that item timing information could be included in
models as collateral information to obtain more accurate estimates of examinee ability. For this use of response latencies though, it would be necessary to know more about response times before they could be incorporated into psychometric models for assessing ability. For instance, if there were conclusive results to show that quicker response rates for correct responses were an indication of higher ability, the response latencies could be included in the final estimates of ability. For example, an item answered correctly in 50 seconds might be weighted in some way and given a higher score than the same item answered correctly in 90 seconds. This is, of course, a greatly simplified example of how timing information might be incorporated into ability estimation. Research by Roskam (in press) addresses this topic in detail.

Finally, another potential use of response times could be in the investigation of differential item functioning (DIF). Item response latencies could be analyzed to determine if certain items are more time consuming for certain subgroups and therefore potentially functioning differentially. Van de Vijver and Helms-Lorenz (1994) define item bias (i.e., DIF) in reaction times as "when individuals with equal abilities or
attitudes from different cultural groups do not have the same expected reaction time for the item." Certain cultural groups may not have had the chance to develop test-taking strategies or may not be as familiar with certain item formats. There could also be cultural differences in the conceptualization of time (Llabre & Froman, 1987) that may affect scores on timed tests. To the extent that any of these are true, this may impact on the validity of their scores (Kendall, 1964).

There have been a number of studies conducted in recent years to address some of the issues mentioned above. Some researchers have conducted exploratory studies to examine what variables might affect response times on computer administered tests. Bergstrom, Gershon and Lunz (1994) used hierarchical linear modeling to investigate the determinants of response time in a computer adaptive test. They found that reaction time increased with item text length and with item difficulty. This seems obvious, however, Schaeffer, Reese, Steffen, McKinley, and Mills (1993) and Kingsbury, Zara and Houser (1993) found a nonlinear relationship between latency and the item difficulty given examinee ability. Kingsbury et al. (1993) showed that there is a slight decrease in
latency as the difference between the difficulty of the item and the estimated ability level increases. It seems that for items further away from an examinee's ability level, the examinee knows relatively quickly either the correct response, or that he or she does not know the answer, in which case the examinee guesses or omits the item.

Bergstrom et al. (1994) also found that response time was higher for items earlier in the test sequence. This result was also found by Kingsbury et al. (1993) who speculated that examinees may get bored toward the end of an exam, and just want to finish as quickly as possible. Another possible reason could be that testing by computer is still a relatively unfamiliar mode of testing. Although examinees are given tutorials to instruct them how to proceed through the test, they may still take more time familiarizing themselves with the computer at the start of a test. A third possible explanation for the slower responses to early items could be that it is a "natural tendency" to begin a test cautiously. Examinees want to do well and so may initially take more time to respond (R. K. Hambleton, personal communication, November 1994).
Also according to Bergstrom et al. (1994), examinee variables, such as age, sex, first language, ethnicity, and estimated ability were not significant predictors of variance in response time. Parshall, Mittelhotz, and Miller (1994) also found that age, sex, and ethnicity showed no relationship with response latencies. The minimal regression effect for demographic characteristics and latency found in these two studies offers encouraging evidence regarding the equity of the computer testing mode.

However, an examination of the relationship between ethnicity, latency and ability shows an interaction effect (Parshall et al., 1994). For some ethnic groups, a higher response latency mean was associated with a lower average test score, for others, a higher response latency was associated with a higher test score mean. To determine how ethnicity and latency may be related, it is important to match examinees from different ethnic groups on ability before comparing item response latencies for the two groups. Once this matching is done, if differences in average item latencies are found, we know that the differences are probably due to bias and not to differences in ability.
Parshall et al. (1994) added that the cognitive classification of the item and correctness of response also showed no relationship with response time. This is opposite to the findings of Bergstrom et al. (1994) and Kingsbury et al. (1993) who showed correctness of response was indeed related to latency. Both research groups found that mean latencies were shorter for correct responses than for incorrect responses.

Work is also being done in identifying aberrant, or unusual, response patterns by studying response latencies. Aberrant response patterns would indicate, for example, examinees who may be cheating or examinees who may have a lack of motivation. Response latencies may also aid in the identification of overexposed items. Kingsbury, Zara, and Houser (1993) developed indices using characteristic correct and incorrect response times that may be useful for identifying aberrant response patterns in computer adaptive testing situations.

Response latency data has also been used in cognitive modeling to enhance the information used in obtaining ability estimates. Tatsuoka and Tatsuoka (1979) were interested in response latencies to aid in the diagnosis and routing of students for different levels of
instruction. For this, they believe it is "necessary to consider an alternative scoring procedure in which individual differences in information-processing skills are taken into account along with individual ability or achievement levels." This is necessary because it is possible for two examinees to have identical responses on a test, but to differ in the cognitive processes used to obtain those responses. Assigning students to the proper level of instruction requires a deeper understanding of how students arrive at responses to particular items, not merely what responses were chosen.

Tatsuoka and Tatsuoka (1979) believe that response latency data can provide clues about the cognitive processing underlying the correct or incorrect responses chosen by an examinee. They found that response latencies follow Weibull distributions. These distributions of response latencies were able to differentiate between students who were taught using different instructional methods. This suggests that the students were probably using different cognitive strategies.

Roskam (in press) also studied response latencies to add to the information used to determine trait level estimates. He described the Rasch-Weibull model to
explain how response time is related to items. This consists of the parameters associated with each of the independent models (the Rasch and the Weibull). The Rasch response time model has mental speed and item difficulty as its parameters; as the time invested in solving an item increases, so does the ability to solve that item. The part of the model based on the Weibull distribution specifies response time as a function of the examinee's mental speed, the persistence of the examinee in trying to solve the item, and the item difficulty. Essentially, mentally faster examinees will take less time to solve an item, the persistent examinees will spend more time, and more difficult items will take more time to solve. These models are complex, and have not yet been widely used in investigations of item response latencies.

Another line of research involves investigating potential item bias in response times. Llabre and Froman (1987) used item response latencies on a computer administered test to compare Hispanic and white examinees. The results of this study were that Hispanic subjects consistently spent more time on test items than white examinees. They also found significant differences between the scores of Hispanic and white students.
Llabre and Froman (1987) claim that the difference between the Hispanic and white students' test scores represents test bias because the subjects were matched on ability. 'Matched on ability' in this study meant that the examinees were all placed in the same level mathematics course. A potential problem exists with this matching, in that the test administered in the study was not a math test; it was the inference subtest of the *California Test of Mental Maturity* (CTMM). Llabre and Froman offer no evidence that the math test used for the matching correlates with the CTMM. Therefore, it is unclear as to whether the students were matched on the ability being measured by the CTMM.

Another explanation for Llabre and Froman's findings is the language differences of the groups investigated. There is a greater chance of Hispanic examinees having Spanish as their primary language instead of English. In fact, this particular sample of Hispanics had spent varying amounts of time in the United States. Some were international students, some were recent immigrants, and many were Cubans. This could definitely have a noticeable effect on response times for items with English verbal content. Given the same test in Spanish, the Hispanic
examinees may have scored as well or better than the white examinees.

Summary

Studies investigating potential differences in testing time for ethnic groups have been conducted for decades at the total test level. The results are inconclusive. Even though timing differences were found in many cases, some studies (Durbin, Osburn & Winick, 1969; Evans & Reilly, 1972; Wild, Durso & Rubin, 1982) suggest that even if the time needed to complete a test does differ for various ethnic groups, allowing more time does not differentially effect the total test score. Therefore, the conclusion often made is that the test is not speeded.

Part of the rationale for declaring these tests (including the LSAT and the GRE) non-speeded is that, for tests with items that increase in difficulty as the test progresses, the examinees that need more time on earlier items most likely would answer the more difficult items incorrectly, even with more time. Therefore, more time would not improve scores of slower examinees.

Lawrence (1993) states that it is not always feasible to order items from easiest to most difficult, as in tests
that contain item sets (e.g., passage-based reading comprehension items). The *Scholastic Achievement Test* (SAT) and *Graduate Management Aptitude Test* (GMAT) are examples of tests with this format, as is the GRE. Because some easier items appear on the test after more difficult and time consuming items, slower examinees often do not reach these easier items because they are spending too much time on the previous, more difficult, items.

Lawrence (1993) was able to show that the SAT and GMAT test forms were more speeded for black examinees than white examinees. This resulted in fewer black examinees reaching the easier items at the end of the exam. Nevertheless, research suggesting that additional time given for a test does not improve scores of one group over than another has been used to defend current test time limits on certain tests.

Now, with the availability of item response times, it is possible to extend the investigation of speededness to the item level. As mentioned above, there are many reasons to study item response latencies; (1) exploratory research is still needed to determine more about the nature of response latencies; (2) response latencies could aid in test construction and test administration; (3)
information about typical response latencies could help identify aberrant response patterns, as in the cases of cheating or item overexposure; (4) item timing could be incorporated into psychometric models in order to provide more accurate ability estimates; and (5) item response latencies could be used in the investigation of differential item functioning.

Because the availability of item response latency information is relatively new, few studies have been done. Much of the work on item response latencies has of yet only been presented at conferences. Very few studies on item response latencies have been published. In the literature search, only one study (Llabre & Froman, 1987) was identified that directly tested for differences in item response times for ethnic groups. The present study adds to this line of research.
CHAPTER 2

PURPOSE OF THE STUDY

The purpose of this study was to investigate response latencies for ethnic groups on a test administered by computer. More specifically, is there a difference in mean response latencies between black and white examinees, matched on ability, for certain item formats? Ethnicity was chosen as a variable because of its importance in all aspects of testing. Every effort must be made to ensure that tests are not biased against any ethnic group. The influence ethnicity may have on response latencies could be an indication of bias. To guard against this potential bias, it is necessary to ensure that examinees with equal abilities from different ethnic groups and different backgrounds have the same expected response latencies for test items.

In this study, black and white examinees were matched on their estimated abilities, as determined by their test score profile. If this matching were not done, any differences found in response latencies could be attributed to group membership, or to different ability levels of the examinees in the two groups. Any differences could either be interpreted as bias or real
differences. Without matching on ability, there is no way to not confound bias and real differences.

There are a number of ways to sort the items in order to study response latencies. Items of different difficulty levels or of different content areas could be investigated. In this study, the items were grouped by item format. This type of sorting was done to determine if there was any particular item type that was more time consuming for one group than another.

The test used in this study was the Graduate Record Examination (GRE) Computer Based Test (CBT). The field test conducted on the GRE-CBT by Educational Testing Service (ETS) did not have the necessary sample sizes to analyze data for ethnic subgroups (Schaeffer et al., 1993). With the data available from the operational GRE-CBT administrations, however, a closer look at response times for black and white examinees was possible.

The response times of the black and white groups, matched on ability, were compared for a number of item types (the item types are described in the Research Method section below). Because the groups were matched on ability, it was expected that their response latencies would also be similar. However, too little is known about
response times, in general, let alone how ethnicity and item type influence response times, to support this hypothesis with existing studies.
CHAPTER 3
RESEARCH METHOD

Examinees

Graduate Record Examination Computer Based Test (GRE-CBT) data for 17,383 examinees who took the test in 1992 and 1993 were made available for this study by Educational Testing Service (ETS). Of the 17,383 examinees, 940 identified themselves as African-American or black. Of these cases, 838 were used in the analyses.

From the 12,866 White examinees, 838 were randomly selected to match the representation of ability levels found in the black group. Ability, for the sake of matching, was determined by the profile of subtest scores (i.e., the combination of the verbal, quantitative and analytical section scores). For example, for each black examinee with verbal score 520, quantitative score 540, and analytical score 530, a white examinee with the same profile of scores was selected. For each black examinee with a given score profile, a matching white examinee was randomly selected from all white examinees who had that same score profile. This was done at all score levels represented by black examinees.
To insure that the examinees used in this study were taking the test seriously, those who spent less than one-third of the allotted section time and those who completed less than one-half of the items in each section were excluded from the analyses. Also, the GRE-CBT scores served as the students' operational GRE scores, therefore the motivation level of these examinees should be equal to that of examinees of any GRE administration. However, potential self-selection factors do exist. First of all, the registration fee for the GRE-CBT was twice as much as the paper-and-pencil GRE registration. Also, those who chose to take the computer administered GRE over the familiar paper-and-pencil format may represent a slightly different population of examinees. These two factors limit the generalizability of any findings.

Graduate Record Examination Computer Based Test

The Graduate Record Examination Computer Based Test (GRE-CBT) is essentially the paper-and-pencil GRE General Test administered by computer. All examinees are administered the GRE-CBT items in the same order as they appear in the paper-and-pencil GRE.

The GRE General Test, and therefore the GRE-CBT, measures certain developed verbal, quantitative, and
analytical abilities that are important for academic achievement....It is not intended to measure inherent intellectual capacity or intelligence. Neither is it intended to measure creativity, motivation, perseverance, or social worth. The test does, however, make it possible to compare students with different backgrounds" (Educational Testing Service, 1994).

For the purposes of this study, data from three different forms of the GRE-CBT were used (forms J, N, and O). Like most large scale assessments, the GRE has many equivalent forms that can be used interchangeably. This is mainly for security purposes, so examinees in the same testing session do not have the same exact test as those seated near them and also so examinees who take the test more than once do not see the exact same exam at each testing. All forms measure the same skills and meet the same specifications for content and difficulty.

The GRE-CBT consists of three sections: verbal, quantitative and analytical.

**Verbal Section**

The verbal measure consists of four types of items: analogies, antonyms, sentence completion, and reading comprehension. Analogy items test the ability to
recognize relationships between two words and to identify a similar relationship between the words in another word pair. An example of an analogy item would be:

1. COLOR:SPECTRUM :: (A) tone:scale (B) sound:waves (C) verse:poem (D) dimension:space (E) cell:organism

Antonym items test vocabulary knowledge and the ability to reason from the definition to its opposite. An example of an antonym item would be:

2. DIFFUSE : (A) concentrate (B) contend (C) imply (D) pretend (E) rebel

Sentence completion items measure the ability to recognize words or phrases that logically complete the meaning of a sentence. An example of a sentence completion item would be:

3. The _____ science of seismology has grown just enough so that the first overly bold theories have been _____.

(A) magnetic..accepted (B) fledgling..refuted (C) tentative..analyzed (D) predictive..protected (E) exploratory..recalled

Reading comprehension items measure the ability to read a passage with understanding, insight, and discrimination. Questions based on the passages deal with explicitly stated points as well as underlying ideas and implications.
of the statements or ideas. An example of a short reading comprehension set would be:

The common belief of some linguists that each language is a perfect vehicle for the thoughts of the nation speaking it is in some ways the exact counterpart of the conviction of the Manchester school of economics that supply and demand will regulate everything for the best. Just as economists were blind to the numerous cases in which the law of supply and demand left actual wants unsatisfied, so also many linguists are deaf to those instances in which the very nature of a language calls forth misunderstandings in everyday conversation, and in which, consequently, a word has to be modified or defined in order to present the idea intended by the speaker: "He took his stick--no, not John's, but **his own**." No language is perfect, and if we admit this truth, we must also admit that it is not unreasonable to investigate the relative merits of different languages or of different details in languages.

4. The primary purpose of the passage is to

(A) analyze an interesting feature of the English language
(B) refute a belief held by some linguists
(C) show that economic theory is relevant to linguistic study
(D) illustrate the confusion that can result from the improper use of language
(E) suggest a way in which languages can be made more nearly perfect
5. In presenting the argument, the author does all of the following EXCEPT

   (A) give an example
   (B) draw a conclusion
   (C) make a generalization
   (D) make a comparison
   (E) present a paradox

Quantitative Section

The quantitative measure consists of two item types: quantitative comparison and problem solving items. In general, the mathematics required does not extend beyond material usually covered in high school. Quantitative comparison questions measure the ability to reason about the relative sizes of two quantities in order to compare them. An example of a quantitative comparison would be:

\[ x + 2y > 8 \]

<table>
<thead>
<tr>
<th>Column A</th>
<th>Column B</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. 2x + 4y</td>
<td>20</td>
</tr>
</tbody>
</table>

A if the quantity in Column A is greater;  
B if the quantity in Column B is greater;  
C if the two quantities are equal;  
D if the relationship cannot be determined from the information given.

Problem solving items are designed to measure basic mathematical skills and understanding of elementary mathematical concepts. An example of a problem solving item would be:
7. If \(3x - 2 = 7\), then \(4x =\)

- (A) 3
- (B) 5
- (C) \( \frac{20}{3} \)
- (D) 9
- (E) 12

Another problem solving item type includes data interpretation items that often occur in sets of two to five items. These items are based on the common information usually presented in a chart or table. An example of this type of item would be:

<table>
<thead>
<tr>
<th>Store</th>
<th>Percent Change in Dollar Amount of Sales in Certain Retail Stores from 1977 to 1979</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>From 1977 to 1978</td>
</tr>
<tr>
<td>P</td>
<td>+10</td>
</tr>
<tr>
<td>Q</td>
<td>-20</td>
</tr>
<tr>
<td>R</td>
<td>+5</td>
</tr>
<tr>
<td>S</td>
<td>-7</td>
</tr>
<tr>
<td>T</td>
<td>+17</td>
</tr>
</tbody>
</table>

8. In 1979, for which of the stores was the dollar amount of sales greater than that of any of the others shown?

- (A) P
- (B) Q
- (C) R
- (D) S
- (E) It cannot be determined from the information given.
Analytical Section

The analytical measure is designed to assess the ability to think logically, both in a rule-constrained and in a common sense way. This section consists of two item types: analytical reasoning and logical reasoning. Analytical reasoning tests the ability to understand the structure of arbitrary relationships, and to be able to deduce relationships from those given. An example of an analytical reasoning item would be:

A farmer plants only five different kinds of vegetables--beans, corn, kale, peas, and squash. Every year the farmer plants exactly three kinds of vegetables according to the following restrictions:

If the farmer plants corn, the farmer also plants beans that year.

If the farmer plants kale one year, the farmer does not plant it the next year.

In any year, the farmer plants no more than one of the vegetables the farmer planted in the previous year.

9. Which of the following is a possible sequence of combinations for the farmer to plant in two successive years?

(A) beans, corn, kale; corn, peas, squash
(B) beans, corn, peas; beans, corn, squash
(C) beans, peas, squash; beans, corn, kale
(D) corn, peas, squash; beans, kale, peas
(E) kale, peas, squash; beans, corn, kale
Logical reasoning items measure the ability to understand, analyze, and evaluate arguments. An example of a logical reasoning item would be:

10. According to one psychological theory, in order to be happy, one must have an intimate relationship with another person. Yet the world’s greatest composers spent most of their time in solitude and had no intimate relationships. So the psychological theory must be wrong.

The conclusion above assumes that

(A) the world’s greatest composers chose to avoid intimate relationships
(B) people who have intimate relationships spend little time in solitude
(C) solitude is necessary for the composition of great music
(D) less well known composers had intimate relationships
(E) the world’s greatest composers were happy

All of the above explanations and examples describing the GRE General Test were taken from the 1994-95 General Test Descriptive Booklet (ETS, 1994).

Procedure

In order to test the question of whether response times are different for black and white examinees on different item formats, the examinees were first matched on ability. As mentioned earlier, ability, for the purpose of this study, was determined by the examinee’s
score profile on the GRE-CBT. Score profile was defined as the scores an examinee received for each of the sections on the GRE.

Due to the large number of possible score combinations, very few perfect matches were found. Therefore, it was necessary to match examinees within a certain range of scaled score points on each section. This matching was done using a computer program written in SIR®. This program cycled through the scores of the black examinees, one at a time, searching for an exact match of scores in the white examinees. If an exact match could not be found, the program would randomly choose to search ten points higher or lower in each section in order to locate a match. The matching process was conducted separately for each of the three test forms used in the study (Forms J, N, and O).

The majority of black examinees (838 of 940) could be matched within 20 points of white examinees on each of the three sections. A range of twenty scaled score points was enough leeway to find matches for most of the examinees without creating large differences in mean section scores. The effect sizes of the difference between black and white examinees (treating the white examinee sample as the base)
are less than .2 for every subtest (see Table 1). In general, effect sizes of .2 or less represent small differences (Cohen, 1988). The table of mean subtest scores on each of the test forms for the matched groups of examinees is shown below.

Table 1
Means and Standard Deviations of GRE-CBT Subtest Scores for Matched Groups of Examinees with Effect Sizes of the Differences Between Groups

<table>
<thead>
<tr>
<th>Form</th>
<th>Subtest</th>
<th>Black Mean</th>
<th>S.D.</th>
<th>White Mean</th>
<th>S.D.</th>
<th>Effect Size (B-W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>Verbal</td>
<td>426.0</td>
<td>103.3</td>
<td>444.2</td>
<td>108.0</td>
<td>-.17</td>
</tr>
<tr>
<td>N</td>
<td>Verbal</td>
<td>437.7</td>
<td>112.1</td>
<td>454.9</td>
<td>120.4</td>
<td>-.14</td>
</tr>
<tr>
<td>O</td>
<td>Verbal</td>
<td>395.7</td>
<td>94.7</td>
<td>410.3</td>
<td>101.9</td>
<td>-.14</td>
</tr>
<tr>
<td>J</td>
<td>Quantitative</td>
<td>426.9</td>
<td>113.0</td>
<td>430.5</td>
<td>104.1</td>
<td>-.03</td>
</tr>
<tr>
<td>N</td>
<td>Quantitative</td>
<td>444.2</td>
<td>124.9</td>
<td>439.0</td>
<td>111.5</td>
<td>.05</td>
</tr>
<tr>
<td>O</td>
<td>Quantitative</td>
<td>393.8</td>
<td>114.2</td>
<td>394.7</td>
<td>94.3</td>
<td>-.01</td>
</tr>
<tr>
<td>J</td>
<td>Analytical</td>
<td>447.1</td>
<td>111.6</td>
<td>432.3</td>
<td>111.4</td>
<td>.13</td>
</tr>
<tr>
<td>N</td>
<td>Analytical</td>
<td>456.8</td>
<td>123.7</td>
<td>448.6</td>
<td>126.9</td>
<td>.06</td>
</tr>
<tr>
<td>O</td>
<td>Analytical</td>
<td>411.3</td>
<td>104.7</td>
<td>400.4</td>
<td>111.9</td>
<td>.10</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>838</td>
<td></td>
<td>838</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The total scores for the examinees were also computed (the sum of the verbal, quantitative, and analytical
subtest scores) in order to categorize the examinees into groups based on overall ability level. Examinees were placed into relatively low (less than or equal to 1140), medium (1150 to 1390) and high (greater than or equal to 1400) ability groups. These categories were created by dividing the matched black and white samples into three groups of equal size.

To provide some perspective on the ability levels of students in this study, a comparison of scores was made to the overall GRE population. For the overall population, the mean total GRE score is around 1500. This can be seen in Table 2 below, taken from Durso, Golub-Smith, Schaeffer, and Steffen's (1994) paper on the comparability of the GRE-CBT and the computer adaptive version of the GRE.

Table 2
Means and Standard Deviations of GRE-CBT Scores

<table>
<thead>
<tr>
<th></th>
<th>Verbal</th>
<th>Quantitative</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Section Score</td>
<td>502</td>
<td>522</td>
<td>538</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>115</td>
<td>131</td>
<td>125</td>
</tr>
</tbody>
</table>

For the group in the present study, 1500 constitutes a score in the high ability category, where 1500 is of
middle ability range for the overall population. It is also evident from comparing Tables 1 and 2 that the group in this study performed less well than the overall population on each subtest. Thus, the comparisons which were carried out between black and white examinees in this study were done so with groups which ranged from relatively low to about average ability in the total population of GRE candidates.

Mean item response time for each of the eight item types described in the GRE-CBT Section above were compared for the three ability levels and two ethnic group categories. Again, the item types are:

Verbal Measure:  
1. Sentence Completion  
2. Analogy  
3. Reading Comprehension  
4. Antonym

Quantitative Measure:  
5. Quantitative Comparison  
6. Problem Solving

Analytical Measure:  
7. Analytical Reasoning  
8. Logical Reasoning

A multivariate analysis of variance (MANOVA) was conducted to test the differences between ethnic groups and ability for the eight item types. MANOVA was chosen
for the analyses because the eight dependent measures (i.e., the item types) are related. For each examinee, a mean response time was computed for each item type. Two levels of ethnicity were crossed with three levels of ability across the eight levels of item type (2 X 3 X 8 design). Multivariate analysis of variance provides an overall significance test that, in effect, tests all possible contrasts.

Additional analyses were conducted to investigate correct and incorrect responses. For these analyses, item response latencies were computed for correct and incorrect responses separately. Two MANOVAs were run to compare black and white examinees for each item type on correct and incorrect responses separately. Also, t-tests were conducted between the means of the correct and incorrect responses to determine if there was a difference in their item response latencies.

In addition, regression analyses were conducted for each of the eight item formats separately for black and white examinees. Regression analyses enable comparisons between the slopes of the groups to determine if the relationship between item response latency and ability was equivalent in both ethnic groups. Ability, based on the
section score corresponding to the item type, was treated as the independent variable in the regression, and response latency was treated as the dependent variable. For example, the verbal score would be treated as the independent variable for the regression on sentence completion, analogy, antonym, and reading comprehension items. Again, Type I error was controlled by using the Bonferroni procedure \((\alpha=.006)\).

As mentioned earlier, there was not enough previous research to support specific hypotheses. However, the prediction was that no differences in response times for the various item types would be found for black and white examinees of comparable abilities. It is because the examinees were matched on ability, that similar mean response latencies were expected. Therefore, any differences that are detected could be associated with problematic item formats.
CHAPTER 4

RESULTS

Main Effect Due to Ethnicity

The MANOVAs detected a significant effect of ethnic group for two of the eight item types: sentence completion (p<.001) and logical reasoning (p<.001). The Bonferroni procedure was used to control Type I error for all contrasts tested, and resulted in an alpha of 0.002. Mean item response times by item type are shown in Table 3.

Table 3
Mean Item Response Times and Standard Deviations for Different Item Types for Black and White Examinees

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Item Response Times</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Black</td>
<td>White</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
</tr>
<tr>
<td>Sentence Completion*</td>
<td>51.13</td>
<td>12.70</td>
<td>47.47</td>
</tr>
<tr>
<td>Analogy</td>
<td>34.32</td>
<td>8.82</td>
<td>34.35</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>78.61</td>
<td>15.53</td>
<td>77.75</td>
</tr>
<tr>
<td>Antonym</td>
<td>20.08</td>
<td>6.10</td>
<td>20.23</td>
</tr>
<tr>
<td>Quantitative Comparison</td>
<td>50.20</td>
<td>13.92</td>
<td>49.96</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>67.73</td>
<td>13.75</td>
<td>67.14</td>
</tr>
<tr>
<td>Analytical Reasoning</td>
<td>76.60</td>
<td>16.65</td>
<td>74.71</td>
</tr>
<tr>
<td>Logical Reasoning*</td>
<td>80.16</td>
<td>20.56</td>
<td>76.69</td>
</tr>
</tbody>
</table>

*Indicates item types with significantly different response times in the black and white samples.
Main Effect Due to Ability

The MANOVAs also detected a significant effect of ability level on response latencies. This was not a surprising finding since it seemed obvious that response time differences would exist across ability groups.

As mentioned earlier, the ability levels are relative to this population of examinees. Again, low ability here is defined as a total score on all three sections being less than or equal to 1140. Middle ability is defined as a total score of 1150 to 1390. High ability is defined as a total score of greater than or equal to 1400.

Differences in response times for ability levels were found in six of the eight item types: sentence completion, reading comprehension, quantitative comparison, problem solving, analytical reasoning and logical reasoning (all at p<.001; see Table 4). For all six item types, there were significant differences between all contrasts of ability levels (i.e., low vs. high, low vs. middle, and middle vs. high).
Table 4
Mean Item Response Times for Different Item Types by Ability Level

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Ability Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low (&lt;=1140)</td>
</tr>
<tr>
<td>Sentence Completion*</td>
<td>53.86</td>
</tr>
<tr>
<td>Analogy</td>
<td>33.88</td>
</tr>
<tr>
<td>Reading Comprehension*</td>
<td>75.37</td>
</tr>
<tr>
<td>Antonym</td>
<td>19.83</td>
</tr>
<tr>
<td>Quantitative Comparison*</td>
<td>47.07</td>
</tr>
<tr>
<td>Problem Solving*</td>
<td>64.13</td>
</tr>
<tr>
<td>Analytical Reasoning*</td>
<td>71.65</td>
</tr>
<tr>
<td>Logical Reasoning*</td>
<td>85.42</td>
</tr>
</tbody>
</table>

N  571  529  576

*Indicates item types with significantly different response times across all ability levels.

A further comparison was conducted on item response latencies by separating correct responses and incorrect responses. Some previous research suggests that examinees spend less time on items answered incorrectly than on items answered correctly (Kingsbury, Zara & Houser, 1993). Speed-accuracy trade-off explanations of item response time also state that less time is often spent on incorrect responses. In this study, however, only three of the
eight item types showed this pattern: problem solving, analytical reasoning and logical reasoning. The other five item types—sentence completion, analogy, reading comprehension, antonym, and quantitative comparison—showed the opposite pattern. For these item types, examinees spent more time on incorrect responses than on correct responses (see Table 5). All differences between mean item response latencies of correct and incorrect responses were significant at less than α=0.006.

Table 5
Mean Item Response Times for Different Item Types by Correctness of Response for Black and White Examinees

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Item Response Time</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Black</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>Black</td>
<td>White</td>
</tr>
<tr>
<td>Sentence Completion</td>
<td>46.55</td>
<td>43.06</td>
<td>61.44</td>
</tr>
<tr>
<td>Analogy</td>
<td>29.93</td>
<td>29.36</td>
<td>38.82</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>77.81</td>
<td>76.46</td>
<td>83.58</td>
</tr>
<tr>
<td>Antonym</td>
<td>17.03</td>
<td>16.92</td>
<td>23.91</td>
</tr>
<tr>
<td>Quantitative Comparison</td>
<td>50.56</td>
<td>49.97</td>
<td>52.46</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>72.01</td>
<td>72.00</td>
<td>66.36</td>
</tr>
<tr>
<td>Analytical Reasoning</td>
<td>90.57</td>
<td>87.78</td>
<td>65.66</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>84.36</td>
<td>79.85</td>
<td>78.01</td>
</tr>
</tbody>
</table>
Separate MANOVAs for correct and incorrect response times found differences between black and white examinees for correct and incorrect responses that were similar to the differences found in the item response times for the combined (correct and incorrect) responses. In both MANOVAs, Type I error was controlled for by using the Bonferroni correction which lowered the alpha of 0.05 by a factor of 8 ($\alpha=0.006$). For correct responses, there was a significant difference between black and white examinees for sentence completion ($p<.001$), analytical reasoning ($p=.004$), and logical reasoning ($p<.001$) items. For incorrect responses, there was only a difference between black and white examinees for response times on sentence completion items ($p<.001$).

**Regression Analyses**

Regression analyses were conducted for mean item response times for each item type across the appropriate GRE section score (verbal, quantitative, or analytical). The regressions were run separately for black and white examinees in order to test the slopes of the two regression equations to see if the slopes were parallel. This was done to determine whether or not the relationship
between item response latency and ability was the same in the two groups.

Plots of the regression lines allow visual comparisons of the black and white examinees. It is also possible to see patterns of response time across score levels. When interpreting the regression lines, the ability levels of the samples used in this study should be kept in mind. The majority of the examinees fall between 250 to 620; relatively few examinees have scores at or above 640 for any of the three sections.

The regression lines illustrate the mean differences between the ethnic groups detected in the MANOVA results, as well as any differences in the slopes. Figures 1 and 2 illustrate the mean differences between ethnic groups for sentence completion and logical reasoning items, respectively. None of the slopes were significantly different.

The MANOVA findings regarding ability level can also be seen in the regression lines. Figures 1 through 6 all show increasing or decreasing trends of item response time across ability level. In order, these figures represent sentence completion, logical reasoning, analytical reasoning, reading comprehension, quantitative comparison,
Figure 1. Response Times of Black and White Examinees Across GRE Verbal Score for Sentence Completion Items
Figure 2. Response Times of Black and White Examinees Across GRE Analytical Score for Logical Reasoning Items
Figure 3. Response Times of Black and White Examinees Across GRE Analytical Score for Analytical Reasoning Items.
and problem solving item types. Notice that there is a positive relationship between response time and ability level for analytical reasoning, reading comprehension, quantitative comparison, and problem solving items, and a negative relationship for sentence completion and logical reasoning items.
CHAPTER 5
DISCUSSION

Black and white examinees were matched on ability, therefore, it was hoped that no group differences in item response latency would be found for the eight item types. For the most part, this was the case. However, some significant effects were detected.

Main Effect Due to Ethnicity

Significant differences in item response times were identified for two item types: sentence completion and logical reasoning. The means in Table 3 show that the significant differences in response time for these two item types translate to differences of 3.66 seconds for sentence completion items, and 3.47 seconds for logical reasoning items. In each case, black examinees spent more time on the items than white examinees. The medians show the same pattern (4.07 seconds difference for sentence completion and 4.05 seconds difference for logical reasoning items). Also, the standard deviations in both groups are about equal (for sentence completion: black=12.70, white=12.97; for logical reasoning: black=20.56, white=20.32).
The first consideration is whether or not these differences are practically significant. The large sample size (N=1676) could have caused the smallest differences to be detected. However, if the large sample size was entirely responsible for the significant differences detected, differences in the other six item types might also be expected.

For sentence completion items, the difference of nearly 4 seconds per item seems small. On average black examinees spend 51 seconds to answer sentence completion items; white examinees spend an average of 47 seconds. When compared to the mean item response times, the four second difference does not seem as if it would cause much of an impact on black examinees' ability to complete sentence completion items in time. With 7 sentence completion items, black examinees would require an additional 28 seconds overall to complete the 32-minute verbal section. Half a minute hardly seems as if it would effect black examinees' ability to complete the verbal section in time.

The logical reasoning items took an average of about 3.5 seconds longer for black examinees to respond. When this is examined within the context of mean item response
times (80 seconds for black examinees, 77 seconds for white examinees), it also seems as though it would have very little effect on black examinees' ability to complete logical reasoning items. In terms of the time difference in relation to section time, with only 6 logical reasoning items per analytical section, an additional 24 seconds is needed overall by black examinees to complete the 32-minute section. Less than one half of a minute overall clearly does not seem to be a large enough time difference to cause concern.

Sentence completion items measure the ability to recognize words or phrases that logically complete the meaning of a sentence. Logical reasoning items measure the ability to understand, analyze, and evaluate arguments. Examples of the two item types were presented earlier.

Differences between black and white examinees for response times on sentence completion and logical reasoning items may not seem practically significant in terms of time needed to complete the items and sections in which these item types are found. However, differences were detected, and these differences need explanation.
Perhaps the reasoning skills necessary to answer these types of questions are not emphasized to the same degree in the schooling or culture of the two ethnic groups. Of course, items are field-tested and routinely analyzed for differential item functioning (DIF) before being included in the GRE. So, the differences in item response times is not due to content DIF in the items. However, the differences in item response times detected in this study could be an indication that these items exhibit item response time DIF (van de Vijver & Helms-Lorenz, 1994). The findings demonstrate the need to ensure that all examinees have sufficient time to complete the test, especially in cases where differences have been found to exist.

Opposite to the null-hypothesis, differences between black and white examinees were detected. No item response time differences were expected because the examinees were matched on their GRE scores. However, the very matching process used in the study could be at least partially responsible for the findings.

The way in which the examinees were matched is not without problems. The black examinees included in the study were representative of the entire population of
black GRE examinees. In fact, with 838 of the total 940 black examinees included in the study, this group nearly was the population. The white examinees were taken from the lower end of the overall population of white GRE examinees in order to be matched with the black examinees.

The examinees were matched on their GRE score profile. However, if this same group were to take the GRE again, the examinees' score profiles would probably be different. Regression to the mean, due to errors of measurement, would pull the white examinees' scores up, while the black examinees' scores would be expected to be about the same. So, although the examinees were matched as accurately as possible in this study, the groups may, in actuality, be more different than they appear.

The shortcomings in the matching could have confounded the results. The MANOVA detected that black examinees spent significantly more time on sentence completion items than white examinees. However, the differences between the two groups on the verbal section means, which include sentence completion items, favored the white examinees. The effect sizes of the differences between the groups in the verbal section were larger than for the other sections. Though these effect sizes are
considered to be small (Cohen, 1988), the fact that black examinees scored slightly lower could account for the slower item response times for black examinees than for white examinees on sentence completion items. As seen in the plot of the regression line of sentence completion items, those with higher verbal section scores did spend less time to answer sentence completion items.

The MANOVA also detected significant differences between black and white examinees in response times for logical reasoning items. In this case though, the differences between the two groups on the analytical section means favored the black examinees. Again, the effect sizes of the differences between black and white examinees on the analytical section scores were considered to be small (Cohen, 1988). As shown in the plot of the regression line of logical reasoning items, those with higher analytical score means took less time to respond to logical reasoning items. If score differences due to problems with the matching were responsible for the significant differences detected for logical reasoning items, the black examinees would be expected to have lower item response times on logical reasoning items. The opposite effect was observed.
Main Effect Due to Ability

A less surprising finding is that differences in response time across ability levels exist. Low, middle and high ability categories had significantly different response times for six of the eight item types: sentence completion, reading comprehension, quantitative comparison, problem solving, analytical reasoning and logical reasoning. It seems obvious that those of different ability would need different amounts of time to respond to test items. What may not be as obvious though, is the direction of the differences. For sentence completion and logical reasoning items, the low ability level needed more time per item than the middle ability examinees, who in turn needed more time per item than the high ability examinees. However, this trend was reversed for reading comprehension, quantitative comparison, problem solving, and analytical reasoning items. For these four item types it was the high ability examinees who spent more time per item.

These findings closely match those found in Schaeffer et al. (1993). The researchers in that study also found that low ability examinees spent more time per item for sentence completion and logical reasoning item types. On
the other hand, high ability examinees spent more time per item for reading comprehension, problem solving, and analytical reasoning items. The present study and Schaeffer et al. (1993) found different results for quantitative comparison items. The present study found that low ability examinees spent less time per quantitative comparison item than high ability examinees; Schaeffer et al. (1993) found the opposite, that low ability examinees spent more time per item than high ability examinees for quantitative comparisons.

The trends found in this study could be explained by a test-taking behavior where lower ability candidates may simply choose to quickly guess a response to items that have a greater amount of stimuli associated with them (e.g., text for reading comprehension and analytical reasoning, data displayed in charts or tables for problem solving, equations or geometry for quantitative comparison). Perhaps this behavior in lower ability examinees is actually a type of test-taking strategy that entails quickly guessing on the item types that have higher textual or stimulus loads in order to have more time to focus on the item types that look less
intimidating (the shorter items, such as sentence completion and logical reasoning items).

When item response times for correctness of response were compared, it was shown that examinees spent more time to answer an item incorrectly for five of the eight item types, and that examinees spent more time to answer an item correctly for the remaining three item types. For some reason, examinees spend more time to answer sentence completion, analogy, reading comprehension, antonym, and quantitative comparison items incorrectly. This trend was reversed for problem solving, analytical reasoning, and logical reasoning items, where examinees spent less time on incorrect responses than correct responses.

It seems that examinees may not be using their time wisely on the five item types where they are spending more time to answer an item incorrectly. Perhaps it is not clear to examinees when they simply do not know the answer to an item, so they continue to try to determine the correct response. Because the GRE is a high stakes test, examinees may refuse to quickly guess or omit items that are too difficult for them, hoping that if they spend time on the item, they will be able to answer it correctly.
The results for the remaining three item types, where examinees spend more time to answer the items correctly may be explained by the speededness of the test. Half of the problem solving items are at the end of the quantitative section. Examinees may be quickly guessing at the end, simply to complete the section. This would result in many quick, incorrect responses.

The analytical section as a whole is somewhat speeded (Schaeffer et al., 1993), so examinees may be quickly guessing responses to both analytical and logical reasoning item types on the second half of the analytical section. Comparing p-values for analytical and logical reasoning items for the first and second half of the section illustrates this point (see Table 6).

Table 6
Comparison of Mean Item Response Times and P-Values for Analytical and Logical Reasoning Items in the First and Second Halves of the Analytical Section

<table>
<thead>
<tr>
<th>Item Type</th>
<th>1st Half of Section</th>
<th>2nd Half of Section</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item Time</td>
<td>p-value</td>
</tr>
<tr>
<td>Analytical</td>
<td>83.23</td>
<td>0.718</td>
</tr>
<tr>
<td>Logical</td>
<td>88.43</td>
<td>0.783</td>
</tr>
</tbody>
</table>
It is clear that examinees are spending more time per item in the first half than in the second half of the analytical section. At the same time, fewer examinees are answering items in the second half of the section correctly.
CHAPTER 6
CONCLUSIONS

It is important and essential to insure that admissions tests are not biased in any way, against any subgroup of the population. This study addressed the potential bias that may exist in the area of item response latencies. By investigating item timing information for different item formats in both black and white examinees, the results of this study will be able to better inform the testing community about item timing factors that may or may not introduce bias into a test.

Differences between black and white examinees were identified in their mean item response times for only two of the eight item types: sentence completion and logical reasoning. Though it is difficult, if not impossible, to determine precisely why these differences exist, it is important to follow-up on this potentially important source of item bias.

It needs to be decided if differences of the magnitude found in this study are large enough to cause concern. The differences in sentence completion and logical reasoning items do not seem practically significant, however, any differences could be potentially
troublesome. These differences in item response times for groups matched on ability could be an indication of differential item functioning (DIF; van de Vijver & Helms-Lorenz, 1994). Perhaps these two item types have some underlying construct in common that is causing the black examinees to spend more time responding than the white examinees.

Future research could investigate the nature of the item types that have differential item response times for different groups to determine if there is some characteristic of the item format that is contributing to the differences in item response times. To follow-up the finding of overall mean response time differences for certain item types, another research direction could investigate mean response times on an item-by-item basis for groups matched on ability to see if certain items exhibit response time DIF. These specific items could then be examined more carefully in an attempt to determine why certain subgroups spend more time than others. It would be interesting to see if items that exhibit response time DIF are also flagged for content DIF.

It may also be useful to interview students or to conduct "talk aloud" experiments (where students talk
about what they are thinking as they take a test) with students of all ability levels about their test-taking strategies. Talking with examinees could help to determine how they allocate their time during a test. Questions could also address how examinees may be affected by different item formats. It would be interesting to compare how examinees of varying ability levels perceive their test-taking experience.

For the present, the findings of this study suggest that when setting time limits, test developers should keep in mind that some subgroups of examinees (e.g., blacks) may require more time to complete sentence completion and logical reasoning item types. The present policy of setting time limits in order for all examinees to have an equal opportunity to complete the test seems sufficient. This should minimize the effects of any score bias due to differential item response times.
BIBLIOGRAPHY


