ERIC
Full Text Provided by ERIC

DOCUMENT RESUME

ED 242 740 TM 840 107

AUTHOR Pike, Lewis W.
TITLE Implicit Guessing Strategies of GRE-Aptitude Examinees Classified by Ethnic Group and Sex.
INSTITUTION Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO GREB-75-10P
PUB DATE Jun 80
NOTE 113p.
AVAILABLE FROM Educational Testing Service, Publications Order Services, Dept I-101, Princeton, NJ 08541.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Aptitude Tests; Black Students; *College Entrance Examinations; *Ethnic Groups; *Guessing (Tests); Higher Education; Hispanic Americans; Racial Differences; Scoring; Scoring Formulas; *Sex Differences; Student Characteristics; Test Bias; Testing Problems; Test Items; White Students
IDENTIFIERS *Graduate Record Examinations

ABSTRACT

This study describes intergroup guessing differences in response to tests and to test-like tasks. It is a composite of seven component inquiries with three substudies in Phase 1 and four in Phase 2. These seven studies cover the Graduate Record Examination (GRE) item-type domain from a number of viewpoints relevant to implicit guessing behavior. The studies in Phase 1 centered on item analytic strategies and on test data derived from a GRE administration. The studies in Phase 2 centered on item-component strategies and on data derived from supplementary materials administered at four university settings. The groups studied were Whites, Chicanos, and Blacks. The implication of any intergroup differences might be that the scoring formula and the instructions to candidates concerning scoring were inappropriate for one or more groups. The most general conclusion is that such intergroup differences do not exist. In seven attempts to find group contrasts, with each attempt yielding a fairly complex and multifaceted analysis, only one minor phenomenon can be reported: Chicano female omitting on GRE-Verbal is demonstrated by individual groups of somewhat lower average ability than that of individuals demonstrating similar behavior for other ethnic-sex groups. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED242740

IMPLICIT GUESSING STRATEGIES OF GRE-APTITUDE EXAMINEES CLASSIFIED BY ETHNIC GROUP AND SEX

Lewis W. Pike

GRE Board Professional Report GREB No. 75-10P

June 1980

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

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 840 107

IMPLICIT GUESSING STRATEGIES OF GRE-APTITUDE EXAMINEES CLASSIFIED BY ETHNIC GROUP AND SEX

Lewis W. Pike

GRE Board Professional Report GREB No. 75-10P

June 1980

Copyright © 1980 by Educational Testing Service. All rights reserved.

Foreword

This report describes a variety of project activities carried out within the framework of a more ambitious initial proposal for a three-stage study of relationships between the guessing behavior of GRE Aptitude examinees, scoring formulas, and within-test guessing instructions. The work of the first stage, concerning the implicit guessing strategies of test-takers, was funded by the GRE Board in September, 1975, and is reported here. The project consisted of seven component investigations with the common theme of a search for ethnic and ability differences in item response behaviors related to guessing. It drew upon the contributions of a number of people; in particular, Phase II depended in large part upon the assistance of interested colleagues at each of four universities who, unfortunately, must remain anonymous.

This final report has been completed and is being submitted to the GRE Board Research Committee after the principal investigator, Lewis W. Pike, left ETS to assume new professional responsibilities. Completion of the report is due in large part to the effort of Thomas F. Donlon, who rearranged and substantially rewrote earlier drafts and prepared significant insertions of his own. Useful suggestions also were offered by a number of reviewers including particularly Robert Altman, Mary Jo Clark, Elsa Rosenthal, Spencer Swinton, and Cheryl Wild. Final copy was prepared by Marian Helms, Christine Sansone, Lorraine Simon, and Sharon Stewart. Without the contributions of all of these people, the report as it stands would not exist.

Executive Summary

The purpose of the present study was to expand the understanding of guessing strategies as these are implicit in the characteristics of responses to test items and components of items by specific subsets of GRE Aptitude test takers. Self-designations "White," "Black," or "Chicano" were used to group examinees; subgroups of White examinees were selected to match the Black and Chicano samples on total GRE score. All analyses were carried out separately for men and women.

The two phases of the present study constitute two methodologically distinct components, each using a different information base. The first involved exploratory analyses of the implicit guessing strategies of the selected subgroups of examinees by means of item-analysis procedures. The data base for this phase was a tape providing the responses of examinees who had taken the GRE Aptitude test in December, 1974. The second phase considered supplementary data gathered specifically for this study to obtain information to expand and interpret the findings from the first phase.

Phase I derived three statistical indicators which could be useful in identifying differences in guessing behavior for ethnic and sex subgroups. The first indicator was an inordinately low level of success on the item, in the sense of proportion passing. This indicator was actually defined in two ways: percent passing of those who reached the item (P+R) and percent passing of those who attempted the item (P+A). An arbitrary level, 16 percent or less passing, was used to identify items with inordinately low levels. These percentages were "inordinate" or "dysfunctional" in the sense that the group could do better, on the average, through unconsidered, random responding. Thus, the indicator identifies items for which "guessing" is not successful for the group.

The second indicator was a mean criterion score (Verbal test score for verbal items, Quantitative test score for quantitative items) that was higher for the group who omitted the item than for those who attempted but missed the item (i.e., those who "guessed" unsuccessfully). This mean for the Omit group was labelled MnO, and an arbitrary level of higher than the 55th percentile for the total group MnO was used to identify items which showed the phenomenon. The logic of the indicator is that more able people may be anticipated to do better, through guessing, than less able people, because they can eliminate more options. When the value of MnO exceeds the overall average, however, more able people are not guessing as frequently as less able people.

The third indicator was relative uncertainty (RU), the Pike and Flaugher modification of the Shannon information index. This indicator reflects evenness or rectangularity of distribution. It has been used in prior work as a general measure of randomness and hence of guessing.

Indices for each item in the GRE Aptitude test, and for subsets of item types, were computed separately for each group of respondents in an ethnic by sex analysis. A detailed inspection of these indices revealed only one finding of a potential difference between groups; this was that Chicano females may omit more readily than other groups. In all other cases, no differences were found in the response patterns of Blacks, Chicanos, and Whites who were matched on total test score. Such differences were found for unmatched random samples of Whites, reflecting their higher test scores. These findings do not simply attest to similarity among the groups in level of success on the items. While the P+R index reflects level of success, the other indices are sensitive to potential differences in the distribution of wrong answers. Their bearing on guessing is inferential, rather than by direct observation, but they do not simply reflect the dichotomy of success or failure on the item.

The results of Phase I, then, clearly support the view that the standard instructions for the GRE Aptitude test are received in similar ways by the various ethnic groups, that the scoring formulas are equally appropriate for these groups, and that there are no differences in guessing behavior independent of differences in average level of ability.

The indicators applied in Phase I are general tests of implicit guessing behaviors. Each derives its relevance for group comparison from a logical relationship which gives it "sensitivity" to some aspect of guessing. Thus, the RU measure, by testing for randomness of response, tests for the presence of uninformed guessing. The P+A and P+R measures, on the other hand, by contrasting score outcomes against chance expectations, test the efficiency of guessing, the extent to which information is effective. Finally, the MnO index, by reflecting the extent to which guessing is associated with higher or lower score, gives some indication of who is guessing.

These general tests lead to inferences of similar item response processes among the groups. However, there is a need to test these inferences further, and Phase II was an attempt to assess the consistency of item reactions across the groups by using tasks conceptualized as analytical subcomponents of solution processes.

Phase II

Phase II consisted of special, exploratory studies of certain ad hoc groups. These groups were selected as convenient samples with which to study the methodological problems. Accordingly, no generalization to populations was possible or intended. In Phase II, there were four special empirical studies of examinee reponses to item components, in an effort to get more direct evidence of the bases for response behavior in several selected item formats. The general spirit of the inquiry was an attempt to find process differences among the groups. While the results of Phase I had indicated general consistency of solution process with respect to the frequency of guessing and the strategies used, Phase II sought a deeper level of analysis. Accordingly, four special measures, based on item components, were developed. Candidate reports with respect to these four special measures were studied for consistency with the results of Phase I and for other inferences.

The four measures were:

1. Word Associations--the subjects rated the strength of association of response words to stimulus words subdivided in analogy items.

2. Contextual Clues--the subjects ranked the appropriateness of answers to sentence completion items, but were given only a short, truncated version of the item stem.

3. Recognition Vocabulary--the subjects indicated which of a set of words they had met before, without being asked to state the meaning or to assert any knowledge.

4. Quantitative Measure--the subjects were given a set of predominantly free response items in very basic mathematical operations, as a sort of mathematical literacy test.

Results in Phase II were generally consistent with Phase I. There were no group differences of any importance.

The common theme of all of these inquiries was the quest for indications of intergroup differences. The implication of such differences might be that the scoring formula and the instructions to candidates concerning scoring might be inappropriate for one or more groups. The most general conclusion of the study is simply that such intergroup differences could not be demonstrated. The seven studies represent a number of attempts to find group contrasts.

ERIC
Full Text Provided by ERIC

Each study was based on a fairly complex and multi-faceted analysis, with several group comparisons. But only one minor phenomenon can be reported: Chicano female omitting on GRE-V is demonstrated by individual groups whose average ability is lower than that of individuals demonstrating similar behavior for other ethnic-sex groups.

While the principle findings are reassuring in terms of bias, the study should serve as an incentive to further work. The impetus here was toward an examination of intergroup differences in item response process. The focus was on guessing process, and the tone and tenor of the study was on the evaluation of the existing program practices in instruction and in scoring. But item process investigations have a valid role of their own; too little is known of item solution process generally, and further work is needed to understand these processes.

Table of Contents

INTRODUCTION

The problem of guessing has long been of concern in testing. Particular attention has been given to the question of whether to discourage guessing, and if so, how, but increasingly a counter-concern has been voiced about the possible need to encourage it. The attitudinal issues involved include the ethics, the scholastic appropriateness, and the efficacy of guessing. Psychometric questions include those of test reliability, validity, and fairness.

Despite the considerable attention given to the problem, relatively little is known about examinee behavior in this area, or about the related issues of how best to instruct examinees about guessing. There are obvious differences in scoring procedures, such as "Rights only" or "Rights-1/_c_ Wrong" (R-W/_c_) (where _c_ is the number of choices minus one). These major variations in scoring technique would have clear relevance for examinee decisions concerning guessing.

Questions regarding behavior which relates to guessing have particular relevance for the GRE Aptitude test. As a comparatively difficult examination, the GRE is perhaps more susceptible to a larger guessing component in the test scores than would otherwise be the case. Another factor is the wide range of candidates served. GRE candidates exhibit more background differences expected among examinees taking any nationally administered test, varying substantially in a number of ways likely to influence their guessing attitudes and behaviors. These include age, level and area of academic preparation, time since last formal education, and amount and recency of experience with standardized tests.

At a time when test fairness is of particular concern, and especially so in regard to the admission to graduate schools of students not in the mainstream of academic preparation for graduate-level education, the question of possible variations in test sophistication or testwiseness (TW) among diverse GRE subpopulations is critical. One aspect of the concept of testwiseness is knowing when to guess, i.e., how to use partial information as a basis for response. Even for the typical GRE examinee, the test necessarily presents many items calling for decisions about answering that must be based on partial information, or even on hunches. This is in part because, as an efficient norm-referenced test, the GRE will necessarily include many items that most examinees cannot readily answer at a level of complete confidence, and in part because by their nature, individual mul'iple-choice questions are not simply all-or-nothing indices of whether something is "known," even though they are scored that way. Underlying an individual's actually having marked the correct answer or not is the probability of his or her doing so, which in turn is a reflection of that person's level of knowledge and reasoning ability vis-a-vis the total item, with its requirement of picking one's way through the several plausible alternative choices.

For the educationally disadvantaged examinee, the problem of guessing may be compounded. First of all, t.ere is likely to be a much higher percentage of test items requiring guessing decisions, and among these, a higher percentage requiring such decisions involving the possible use of hunches, rather than of firm, though partial,' information. Secondly, the educationally disadvantaged examinee is likely to be relatively deficient in testwiseness, and thus less likely to know how best to use the information and reasoning powers at his or her disposal. Thus, the minority candidate is potentially exposed to a kind of double jeopardy, with weaknesses in developed verbal and quantitative reasoning abilities compounded by problems in coping with the test *per se*.

The implications for fairness are obvious. Differences in test performance between groups should not depend upon such secondary factors as differences in test wiseness. The goal of an equitable testing program must be that of identifying and using the best combination of scoring procedure and within-test guessing instructions related to that procedure. But little is known concerning actual candidate behavior. Guessing behaviors must be inferred from test outcomes. While the typical item analysis will shed some light on the problem, more along these lines can be done. This research was formulated as a study of the guessing strategies which are implicit in the statistical outcomes of test items and of special item-component tasks. Before describing the study, however, a conceptual framework is offered.

Conceptual Framework

It is useful, in presenting the present study, to briefly review its basic assumptions in three primary areas relating to guessing: 1) that there are levels of information upon which guessing is based 2) that guessing behavior is ethically appropriate in the measurement context and 3) that the psychometric effects of guessing behavior on scores are potentially practical and useful.

Levels of information. There are three general levels of information with which an examinee may confront a test item:

full information (FI), partial information (PI), and no information (NI). The FI and NI situations are essentially straightforward in terms of their behavioral consequences, but the PI situation is more complex and more interesting. A useful approach to item-PI is to distinguish among full, partial, and no information at the choice level. This produces complex situations. For example, a special case of item-PI which is often implicit in discussions of guessing behavior is one involving some full information at the individual choice level. That is, one or more distractors may be fully known to be wrong, allowing for an information-based elimination of these choices from consideration.

A consideration of the different levels and kinds of information involved in answering multiple-choice test items is central to the study of guessing behavior. (Here, and in subsequent discussion, "information" will be used in a generic sense to include "comprehension," "computation," reasoning," and so on, as required for answering test items.) As noted above, a candidate may confront a test item at one of three perceived levels of information: full information (FI), partial information (PI), or no information (NI). Although the FI and NI item situations are essentially straightforward, it should be noted that true NI is probably much less common than is usually assumed. A corollary is that truly "blind" guessing, so often cited as a major concern, is almost surely rare. Instead, the common alternative to fully informed guessing is most likely to be guessing on the basis of vague hunches or misinformation. This fact has clear implications regarding guessing formulas and instructions, some of which are explored in later sections of this report.

Attitudinal questions. Whether guessing is considered ethical or scholastically appropriate has a direct bearing on the guessing behavior of examinees and on the positions taken by educators to influence such behavior. The scholastic appropriateness of using PI is best defended by considering that this is:

(1) Consistent with the generally accepted psychometric assumption that the knowledge or ability underlying the essentially discontinuous multiple-choice item format is in fact a continuous variable.

(2) Consistent with concepts of educational outcomes as broader shifts in behavior, rather than simple acquisitions of all-or-none mastery of previously defined content.

ERIC
Full Text Provided by ERIC

Beyond scholastic appropriateness, the additional questions of whether it is ethical or worthwhile to guess can best be answered by considering them simultaneously. Examinees sometimes reason that if they receive full credit on a PI item, because they guess correctly, they thus have an unfair advantage. On the other hand, if they guess incorrectly, they will receive a "deserved" penalty. By this reasoning, either outcome is perceived as an indication that one should not guess: one is "wrong" and the other leads to a lower score. The questions can be resolved simultaneously, however, by demonstrating that over a set of PI items, the expected outcome of using partial information is receiving partial credit, a result that can be recognized as both fair and efficacious.

Psychometric questions. For the student, the basis for decisions about the fairness of guessing is in large part attitudinal. For the sponsors of testing, the question of fairness associated with guessing is psychometric, as well. Differences in test scores attributable to individual guessing tendencies constitute a source of unfairness, whether due to differences in risk-taking tendencies (deciding when to guess) or in test sophistication (knowing when and how to guess).

A common view of guessing behavior by measurement workers considers it a component of the overall score but as not content-related. Cureton, for example, suggests that for multiple-choice tests, "....the true score is the true content score plus the true guessing-tendency score (1971, p. 829)." It is the position of this report that partial information about the item, particularly in the form of choice PI, is itself part of the true content score, and that efforts to (1) encourage proper use of PI, (2) discourage guessing in NI situations, and (3) discourage the use of PI when it is in the form of vague hunches will result in a more consistent matching of guessing behavior to the examinee's level of information with regard to particular items and item choices. Such matching will, across examinees, maximize the valid component of guessing, while minimizing the spurious component--i.e., individual differences in guessing tendencies not directly related to the underlying information or ability that is of interest.

The effect of guessing behavior on score reliability is most evident when PI and NI are considered in terms of the number of distractors that can be eliminated. When examinees do not guess, those who are able correctly to eliminate one, two, or three distractors from a given item are indistinguishable from the NI examinees: all receive a zero item-score, leaving no basis for discrimination among the four levels of content information. If such examinees do guess, however, the differences in their levels of item information are reflected in expected item score differences. This added true score variance will, of course, yield added score reliability. Guessing in NI situations, on the othe hand, will clearly reduce reliability, since it will contribute error variance only.

Guessing involving PI in the form of vague hunches is more complex. It may or may not add to reliability. When highly unsystematic, it will, like NI, reduce reliability, but when it is highly systematic (whether or not it is correct), it could well increase reliability. In some instances the contribution of such guessing to reliability will be positive due to the systematization introduced by particularly compelling distractors. In those instances, examinees will have less than chance success in answering the item correctly, and encouraging such guessing could mean gaining reliability at the expense of reduced fairness and validity.

In general, from the psychometric standpoint, the use of PI is appropriate. The measurement person seeks to provide a mutual-benefit matching between an examinee's wish to make the most of his or her information (and therefore putting PI to use by guessing), and the goal of optimizing test reliability and validity.

In examining the effects of scoring formulas and guessing instructions on guessing behavior, the possibility of systematic differences in guessing associated with examinee characteristics should also be examined. Individual and group differences which influence guessing include attitudes toward risk-taking, and toward the ethical and scholastic legitimacy of guessing, and levels of understanding of how to adopt an optimal guessing strategy to match the scoring procedure. One difficult psychometric area concerns differences in guessing associated with item format. These include differences in (1) the basis for answering: e.g., Vocabulary items call primarily for information; Reading Comprehension items involve comprehension, inference, and locating information; Analogies require verbal reasoning and knowledge of the subtleties of word meaning; and Quantitative items test a combination of knowledge, computation, and mathematical reasoning, and (2) whether the correct answer can be selected independently of the other choices (it can in many antonyms, for example, but cannot in items such as "Which of the following values is greatest?"). Analogies tend toward the "best answer" end of this scale. Other item characteristics, such as whether the choices call for making fine distinctions, cut across item format categories. For each of these item characteristics there can be different kinds of PI, with associated differences in optimal guessing strategy.

STUDY DESIGN

Purpose

The purpose of the present study has been to expand the understanding of guessing strategies as these are implicit in the characteristics of responses by groups of test-takers to individual test items and to item-component tasks. Through the development of such expanded understandings, it may be possible to evolve techniques for test instructions and scoring which are optimally appropriate to the needs of a number of diverse groups. The information obtained in the current study was also seen as having implications for test specifications and test development, and as offering useful insights into the basic processes by which examinees answer multiple-choice questions of various kinds.

The study may be conceived of as having two phases. Phase I consisted of an analysis of a data tape providing the responses of examinees who took the December 1974 GRE aptitude test. This data base in addition to the large number of examinees represented, included responses to a background inventory providing such information as ethnic group membership, sex, and major field of study. Thus, it was possible to explore both examinee and item characteristics as these related to indications of systematic differences in guessing behavior. In Phase II, there were special empirical studies of examinee responses to item components, in an effort to get more direct evidence of the bases for guessing behavior in several selected item formats. This phase developed information by administering supplementary materials to four groups of college students.

Two major limitations inherent in the design of the present study were recognized at the outset. First, the data reported here in the Phase I analysis are based on only one scoring technique, R-W/_c_, and on the standard guessing instructions which accompany this technique. Second, the supplementary data in Phase II were obtained from subjects other than those taking the 1974 GRE, so that the linkage between the findings in the two phases is indirect.

Recognizing these limits, the project proposal clearly indicated that the present study should be considered preliminary to additional research which would examine differentially the effects on guessing behavior of three scoring formulas: R, R-W/_c_ (the "penalty for guessing" adjustment), and R + Omits/_n_, where _n_ is the number of alternatives (a "reward for not guessing" correction). In addition, differences between the standard guessing instructions now in use, and expanded instructions, would be examined. Such further research would explore a two-stage guessing model, in which (1) some distractors are eliminated on the basis of full information at the individual option level, and (2) a non-eliminated alternative is selected as the answer to the item, as these relate to the different scoring formulas and guessing instructions. Further, the experimental subjects who would provide the two-stage guessing model responses would be the same persons responding to the test materials, thus allowing analyses directly linking test performance and guessing sponding to the test materials, thus allowing analyses directly linking test performance and guessing strategy in a manner not possible within the framework of the current effort.

Despite its limitations, however, the present study was recognized as a useful exploratory effort adequate to serve three basic purposes. The first was to test and refine certain tentative hypotheses regarding differences in guessing behavior associated with various combinations of examinee and item characteristics. The study sought to document instances in which some examinees may be systematically at a disadvantage on certain classes of items, in that they tend to guess or omit in ways that yield lower expected scores than are received by examinees who use information more effectively. The second purpose of the study, which was primarily exploratory rather than confirmatory, was to search for ways of counselling examinees when and how to guess in situations where they hold partial information (PI). Finally, a third purpose was to serve as a feasibility check for some of the later work proposed and as a basis for designing and carrying out such work.

Methods

The two phases of the present study constitute two methodologically distinct components, each using a different information base. The first involved exploratory analyses of the implicit guessing strategies of selected subgroups of GRE Aptitude examinees, by means of what are essentially item-analysis procedures. The data base for this phase is a tape providing the responses of examinees who had taken the GRE Aptitude test in December, 1974. The second phase made use of supplementary data gathered specifically to obtain information needed to expand and interpret the findings stemming from the first phase.

Because of the distinctiveness of these phases, and because each involves relatively detailed presentations of complex and somewhat unfamiliar procedures, the report is organized into two distinct sections, presenting the methods and results for each phase separately, before returning to a joint discussion in the final section.

PHASE I

The GRE Aptitude Test administered in December, 1974 included in Section I 55 verbal discrete items: 18 analogies, 20 antonyms, and 17 sentence completion items. Section II was made up of 40 reading comprehension items, based on six reading passages, and Section III contained 55 quantitative items. This form of the GRE also included a set of background questions including ethnic group membership, undergraduate major field, planned graduate major field, and current educational status.

The population of examinees from which samples were drawn consisted of the 70,888 candidates. This number was reduced by excluding examinees who indicated that they did not communicate best in English, and was further limited to those indicating either "college senior" or "college graduate" as their current educational status, and indicating a graduate degree objective of masters, intermediate, or doctorate. After these exclusions, the numbers of males and females for three ethnic groups selected for study were as follows: Caucasians, 15,362 and 13,754; Chicanos, 195 and 130; and Blacks, 915 and 1715. From each Caucasian group, a random sample of 2000 examinees was then drawn for subsequent item analyses. The other groups were used in toto.

Means and standard deviations of GRE-V and GRE-Q scores for these examinee groups are given in Table 1. Formula scores on the 95-item GRE-V for the White, Chicano and Black groups were approximately 49, 34, and 25 points, with a standard deviation for each group of about 16. There were only slight differences by sex within each ethnic group. The pattern of formula scores on the 55-item GRE-Q was similar, with means of about 30, 20, and 14, and standard deviations of about 10. Mean score differences by sex were observed for the White, Chicano and Black samples, with males leading females by about .75, .35, and .50 standard deviations, respectively. In order to examine whether differences in guessing behavior associated with ethnic group and sex could be attributed to differences in overall ability as measured by the GRE, matched groups of whites within sex were drawn for each minority group. The matching was based on combined GRE-V scores and GRE-Q scores, that is, the simple sum of the GRE-V and Q. Summary data for these examinees are also given in Table 1. For the male matched groups the GRE-V means are slightly lower, and the GRE-Q means slightly higher, than those observed for the original minority samples. These differences were not observed between original and matched female groups.

While the matching operation is a useful approach to reducing intergroup differences attributable to differences in level of ability, it cannot totally remove such factors. In the current study in which Whites are selected for the matched group on the bases of scores lower than the mean, the average "true score" of the matched group will be higher than that of the actual ethnic or race sample. This is so because in regression theory the matched group is likely to contain more negative errors of measurement. It is a variation on the familiar regression toward the mean. These known

Table 1

Means and Standard Deviations of GRE-V and GRE-Q Formula Scores of Selected Original and Matched-Sample Examinee Groups

Examinee Group		N	GRE-V (95 items)		GRE-Q (55 items)	
			Mean	SD	Mean	SD
			Original Groups			
Males	White	2000[a]	48.6	15.8	33.3	10.4
	Chicano	195	33.9	15.5	21.4	11.0
	Black	915	25.5	16.8	16.7	10.7
Females	White	2000[a]	49.2	16.2	26.1	9.1
	Chicano	130	33.2	17.9	18.2	9.4
	Black	1715	23.5	15.4	12.2	8.7
		Matched-Sample Groups[b]				
(White males)	Chicano	570	32.8	15.5	24.3	11.1
	Black	350	23.2	15.4	17.0	11.4
(White females)	Chicano	390	33.4	16.4	18.3	9.7
	Black	360	24.7	15.4	13.2	8.3

[a] White male and female groups were randomly selected from pools of 14,733 and 13,132 examinees, respectively.

[b] Matched samples of Whites within sex for each minority group were drawn, based on combined GRE-V and GRE-Q score. For Chicano males, two Whites were drawn for each Hispanic male (195 Chicano and 90 Puerto Rican) in the original sample, at each score level. (The two Hispanic male groups had nearly identical score distribution patterns. Compare male Puerto Rican means and standard deviations of 34.0 and 18.0 for GRE-V, and 21.3 and 12.1 for GRE-Q, to data for male Chicanos given above.) For Black males, Chicano females, and Black females, the selection ratios were, respectively, one, three, and two.

difficulties in matching are generally of a tolerable magnitude; it was similarly judged that the errors of matching due to the use of combined GRE-V and GRE-Q scores were tolerable, and that the convenience which the combination afforded outweighed its disadvantages.

Indices of Guessing Behavior

Three kinds of item data were evaluated as indicators of aspects of guessing behavior. These indicators were seen as sensitive to the use of partial item information, such as the use of full choice-level information to eliminate one or more distractors, and as reflecting the avoidance of hunch-based guessing where this reduces the expected item score under the formula score (R-W/c) condition. Two of the three kinds of item information, percent-pass and omitting patterns, are provided in standard ETS item analyses. The third, relative uncertainty (RU), is not routinely computed; it indicates the degree to which error responses for an item are either relatively concentrated on three or fewer of the four error choices, or relatively evenly spread over all four. The role of these indices in explicating examinee guessing is discussed in the following section.

Low percent-pass (Low P+). The percent-pass (P+) value for a given test item that is observed for a given examinee group reflects some composite of the full information, partial information, and random-like guessing being used by the examinees in that group. But the strength of these different levels cannot be determined. For this reason, differences in P+ between examinee groups are not ordinarily informative with respect to possible differences in guessing strategies. However, occurrences of P+ values clearly below the chance level of 20 (for 5-choice items) provide direct evidence that optimal guessing strategies are not being used successfully.

A distinction must be made at this point between P+R, which uses the number of examinees reaching a given item as the denominator, and P+A, which uses the number of examinees answering the item, i.e., making a marked response.

The first of these, P+R, is the value widely used as an index of item difficulty. Low P+R values may indicate items for which an examinee group had less success than if the entire group had answered the item in a random-like way. As P+R values approach zero, the likelihood of a significant amount of random-like guessing decreases. However, the amount and the appropriateness (success rate) of guessing remains generally indeterminant for a number of levels of P+R. A P+R of 10, for example, would be the expected value in each of the following situations: (a) If 10 percent knew the answer and marked it correctly, and the other 90 percent omitted the item; (b) If 5 percent knew and answered correctly, 25 percent selected randomly among the 5 choices, and 70 percent omitted; (c) If 50 percent answered randomly, and the other 50 percent omitted; and (d) If 10 percent knew and answered correctly, and the other 90 percent all answered incorrectly.

The second percent-pass statistic, P+A, gives the percentage of examinees actually answering a test item who marked the correct choice, and as such, directly indicates the success-rate of those answering the item. For the four instances just noted, each yielding an expected P+R of 10, the expected P+A values would be 100, 33, 20, and 10, respectively. When the P+A value is clearly less than chance, then, it may be inferred that optimal guessing strategies are not being used successfully. This inference does not rule out the possible validity of test items for examinee groups experiencing low P+A values on them. However, it does suggest a strong possibility that some of the score variance attributable to such items is due to a failure to use optimal guessing strategies, rather than to the verbal or quantitative reasoning ability the items are intended to measure.

For the present analyses, instances of $P+R \leq 16$ and $P+A \leq 16$ were arbitrarily classified as low P+, and patterns of such instances associated with examinee groups and item formats were examined for implications regarding guessing behavior. Where such low values are exhibited, the guessing decisions are clearly worse than chance.

Reverse omitting pattern. One way to juo the appropriateness of a group's omitting behavior on a given test item is to examine the pattern of omissions across increasing levels of ability within the group. The usual pattern, when examinees are properly judging when to guess, shows a decreasing rate of omitting accompanying increasing levels of ability. This assumes that under conditions of uncertainty the candidates' options are to guess or to omit, and that guessing should be more appropriate as level of ability increases. A reversal of this pattern of omitting behavior was observed by Flaugher and Pike (1970), for responses of inner-city Blacks on the PSAT-V and PSAT-M. Similar reverse patterns have been observed by Echternacht, Carlson and Flaugher (1973) for minority UP-V and UP-Q responses, and by Conrad and Wallmark (1975) for responses to the GRE-Q.

A pattern of Reverse Omitting, in which the higher scoring members of a group more often omit a given item than do the lower scoring ones, is indicated by a mean test score for the omitting subgroup that is higher than the overall group mean. In this study, the item analysis program assigned criterion scores to each group with an arbitrary mean of 13.0 and a standard deviation of 4.0. In the study, the value of 13.5 was arbitrarily selected as the critical level of the mean score for these omitting. When MnO = 13.0, the group averages about the 50th percentile. When MnO = 13.5

the group averages about the 55th percentile; when MnO = 14.0, the group averages about the 60th percentile.

Relative uncertainty (RU). This measure, based on Shannon's (1949) Uncertainty measure, was applied by Pike and Flaugher (1970) as an index of the relative amounts of randomness or spuriousness in different examinee groups' error responses to a given item. The procedure was used to compare the PSAT-V responses of a group of inner-city Blacks to those of a standard reference population. Error responses of the inner-city group were significantly more random-like than those of the reference group for Sentence Completion, Reading Comprehension, and Antonym items, but not for Analogies. Applying similar procedures to UP-V and UP-Q data, Echternacht, et al. (1973) found significant differences in the randomness of Black and White examinees' responses to Antonym, Reading Comprehension, and Quantitative items, but not to Sentence Completion and Analogy items.

Shannon's index of uncertainty was defined as:

$$U = - \sum_{i=1}^{K} P_i \log_2 P_i$$

where K = the number of categories in a distribution and Pi = the proportion of the distribution in a category.

Shannon's measure has a variable maximum depending upon the number of categories. To adjust for this, Pike and Flaugher proposed relative uncertainty RU, defined as:

$$RU = \frac{- \sum_{i=1}^{K} P_i \log_2 P_i}{K}$$

To the extent that a distribution is rectangular, RU values approach 1.0. In a sense, the RU index is a convenient, informal alternative to a Chi-square test of the property of rectangularity of distribution.

The general logic of these three indicators derives from a model which views informed guessing as a worthwhile strategy, blind guessing as more or less neutral, but which considers th. some "guessing" is probabilistic responding based on misinformation. In this context, the three indices have somewhat different functions. The RU index will reflect when blind guessing is present; the MnO index will reflect who is guessing (by inference from the ability of those omitting); the P+ indices will reflect the presence of dysfunctional responding, lower than chance expectation, and presumably based on misinformation.

None of these indices is direct, in the sense of reflecting observed guessing. Each is inferential, presumed to reflect guessing behavior where this is defined not only as blind responding in the absence of no information but as responding in the context of partial information or misinformation. The measures have in common a ready derivability from routine item analysis.

Findings for Items Grouped by Format

The study of the data-tape information for indications of guessing behavior proceeded from the general to the specific, looking first at larger classes of items and then at smaller ones. In the report, summary data for items grouped on *a priori* bases will be examined first, followed by data for individual items.

Findings regarding each of the three indices related to guessing behavior (Low P+, Reverse Omitting, and Relative Uncertainty) will be considered in order. Each presentation considers the items grouped by format. Among GRE-V items, these formats are Analogies, Antonyms, Sentence Completion, and Reading Comprehension. For GRE-Q, items are divided into two groups, Data Interpretation items and "Other Quantitative" items. Data Interpretation items are presented to examinees in clusters; all items in the cluster refer to information presented in an associated table or graph. Each "Other Quantitative" item is presented independently.

Low P+R and Low P+A. Frequencies of Low P+R items (based on percentages of examinees reaching an item) and of Low P+A items (based on percentages of examinees who answered the item) are given in Table 2 for male examinees and Table 3 for female examinees with subjects grouped by ethnic status and items are grouped by

Table 2

Numbers of Items within Item-Format Answered with Low $P+_R$ or Low $P+_A$

by Original and Matched-Sample Male Examinee Groups

		Male Examinee Group					
		White		Chicano		Black	
Item Format	Number of Items	Low $P+_R$	Low $P+_A$	Low $P+_R$	Low $P+_A$	Low $P+_R$	Low $P+_A$
		Original Groups					
Verbal							
Analogies	18	5	3	6	4	9	6
Antonyms	20	2	0	5	1	5	1
Sent. Completion	17	0	0	1	0	1	1
Rdg. Comprehension	40	0	0	0	0	0	0
Quantitative							
Data Interpretation	14	0	0	2	1	4	1
Other[a]	41	1	0	2	2	2	2
(No. of examinees)		(2000)		(195)		(915)	
	Matched-Sample Groups			Whites matched to Chicanos		Whites matched to Blacks	
Verbal							
Analogies	18			7	5	8	5
Antonyms	20			6	1	6	1
Sent. Completion	17			1	0	3	0
Rdg. Comprehension	40			0	0	0	0
Quantitative							
Data Interpretation	14			2	2	5	3
Other[a]	41			1	1	3	2
(No. of examinees)				(570)		(350)	

Note: Low $P+_r$ is defined as percent-pass $\leq .16$, computed with the number of examinees reaching an item as the base N; Low $P+_a$ refers to percent pass $\leq .16$, computed with the number of examinees actually answering an item as the base N.

[a] "Other" includes 10 algebra, 14 geometry, 12 arithmetic, and 5 miscellaneous items.

Table 3

Numbers of Items within Item-Format Answered with Low P+ by Original and Matched-Sample Female Examinee Groups

		Female Examinee Group					
		White		Chicano		Black	
Item Format	Number of Items	Low $P+_R$	Low $P+_A$	Low $P+_R$	Low $P+_A$	Low $P+_R$	Low $P+_A$
Original Groups							
Verbal							
Analogies	18	4	3	6	4	9	5
Antonyms	20	1	0	3	0	7	1
Sent. Completion	17	0	0	1	0	1	1
Rdg. Comprehension	40	0	0	0	0	0	0
Quantitative							
Data Interpretation	14	2	1	5	5	6	5
Other	41	2	1	2	2	6	4
(No. of examinees)		(2000)		(130)		(1715)	
Matched-Sample Groups							
Verbal							
Analogies	18			7	4	9	5
Antonyms	20			6	1	7	1
Sent. Completion	17			1	0	2	1
Rdg. Comprehension	40			0	0	0	0
Quantitative							
Data Interpretation	14			5	4	6	5
Other	41			3	1	6	3
(No. of examinees)				(390)		(360)	

format. The frequency of Low P+R items was clearly related to ethnic group membership. Of the 95 GRE-V items, 7 Low P+R's were observed for White males, 12 for Chicano males, and 15 for Black males; for the 55 GRE-Q items, there were 1, 4, and 6 low P+R's for White, Chicano, and Black males. These differences are consistent with differences among the respective examinee groups in GRE-V and GRE-Q mean scores, since lower mean scores should be coincident with a greater number of items being very difficult for a given examinee group. For White males matched to Chicano and Black males on the basis of combined GRE-V and GRE-Q scores, the frequencies of Low P+R's departed only slightly from those for the original minority groups, and then in a direction consistent with the fact that t e White matching groups had slightly lower verbal scores and slightly higher quantitative ones than was true of the minority groups. Thus, differences in the frequency of Low P+R items are consistent with differences in overall test difficulty for all groups. This suggests that there is no reasonable basis for attributing observed P+R differences directly to ethnicity.

The findings for females are analogous: frequencies of P+R are 5, 10, and 17 for White, Chicano, and Black females on the Verbal test; 4, 7, and 12 on the Quantitative. Again, the matched White group performs similarly to the related ethnic group. There is little evidence of ethnic linked differences on this index of inefficient guessing.

When the separate item formats are considered, it becomes evident that there are strong differences. Among GRE-V items, most instances of Low P+R occurred for Analogies and Antonyms; there were none for the 40 Reading Comprehension items, for either sex. Among GRE-Q items, there were about as many Low P+R's among the 14 Data Interpretation questions as among the other 41. There is apparently something in the nature of the Data Interpretation items which makes them more difficult to guess on.

The frequencies of Low P+A values are subject to a similar analysis regarding possible guessing strategies, since each such instance indicates an item for which those members of a given examinee group actually answering the item did so with less than chance-level success. Among the GRE-V items, only about half of those in the Low P+R range were also in the Low P+A range. The observed P+A's followed the same general pattern across examinee groups as noted for observed P+R's (i.e., a pattern consistent with the groups' mean GRE-V scores). When the separate item formats are considered, however, the Low P+A pattern departs from that for Low P+R's. For Analogies, considering Tables 2 and 3, 25 out of 39 instances of low P+R were also instances of low P+A, but only 3 out of 24 of the low P+R's for Antonyms were also instances of low P+A. This would suggest that there was less omitting and more dysfunctional guessing in response to very difficult Analogies than to equally difficult Antonyms. Among the GRE-Q items, there was only slight attrition between instances of Low P+R and of Low P+A, suggesting that for these items, as for Analogies, there exists a tendency for very hard items to be answered at a worse than chance level (i.e., for "dysfunctional guessing").

The MnO index. The number of instances in which the omitting subgroup's mean percentile rank was at least 55 (MnO $\geq$ 13.5) is given in Table 4. Among the GRE-V items, instances of reverse Omitting were concentrated among Blacks, both male and female, primarily within the Analogies and Antonyms. Approximately half the Analogy items show the Reverse Omit phenomenon for Black males, and about a third of these items for Black females. Among Data Interpretation items in the GRE-Q, the largest number of item Reverse Omit was again demonstrated by Blacks, but the frequency for Chicanos was nearly as great as that for Blacks on the Other Quantitative items. Patterns of occurrence of Reverse Omitting by the four White groups that were matched to the four minority groups closely resembled those of the original minority groups, with Whites matched to Blacks showing the Reverse Omit pattern primarily on Analogies, Antonyms, and on Other Quantitative items, and with those matched to Chicanos doing so primarily on Other Quantitative items.

As was true of observed differences in P+ patterns, then, the most parsimonious interpretaton of Reverse Omitting differences between sexes and ethnic groups would be to attribute these differences to the observed differences in the relevant tested abilities, rather than to sex or ethnicity, in themselves. On the surface, however, the finding clearly suggests that white males matched in ability to the Black sample show higher incidence of "dysfunctional omitting" than their Black counterparts.

Relative uncertainty. The utility of the RU statistic is that for a given item and for a given examinee group, it summarizes the dispersion over the several error categories. The information is useful, because it allows tentative inferences regarding the degree to which the error responses are either highly systematic or essentially spurious or random-like. The maximum RU of 1.00 is

Table 4

Number of Items within Item-Format Answered with Reverse Omitting by Original and Matched-Sample Examinee Groups

		Examinee Group					
	Number of	Male			Female		
Item Format	Items	Wh	Ch	Bl	Wh	Ch	Bl
		Original Groups					
Verbal							
Analogies	18	0	1	10	0	0	7
Antonyms	20	0	2	7	0	0	4
Sent. Completion	17	0	0	2	0	0	1
Rdg. Comprehension	40	0	0	2	0	0	2
Quantitative							
Data Interpretation	14	0	0	1	0	0	2
Other Quantitative	41	0	3	5	1	5	7
(No. of examinees)		(2000)	(195)	(915)	(2000)	(130)	(1715)
		Matched-Sample Groups					
Verbal							
Analogies	18		1	11		1	3
Antonyms	20		2	11		1	3
Sent. Completion	17		0	0		0	0
Rdg. Comprehension	40		0	3		2	4
Quantitative							
Data Interpretation	14		0	0		0	1
Other Quantitative	41		1	3		3	8
(No. of examinees)			(570)	(350)		(390)	(360)

Note: The criterion for classifying an omitting pattern as Reverse Omitting is that the standardized mean score of examinees within a given group equals or exceeds 13.5. This corresponds to a group mean percentile on the test of 55. I.e., the better students in the group are more often omitting than are the poorer students.

ambiguous, in the sense that it can occur for either completely random responding, or for a special case of fully systematic responding, in which each of the error choices appeals to a different subgroup of examinees, with the subgroups happening to be of equal size. As RU approaches its minimum value of zero, however, there is increasing evidence of systematic error selection.

The RU index is a very general appraisal, however, and its item evidence of systematic error is ambiguous with respect to a basic differentiation in answering strategies suggested in the present paper. That is, answering on the basis of successive eliminations on the one hand and answering primarily on the basis of being drawn to the most plausible or attractive alternative, on the other. Some insight can perhaps be achieved by comparing RU values on a given item, for groups differing in mean total score. Where two groups differ substantially in average score, it is reasonable to expect that on most items, the lower scoring group will be able to eliminate fewer alternatives on the basis of choice-level full information. It follows that the responses of this group would be more generally dispersed over the error categories than would be true of the higher scoring group and that they will demonstrate higher values of RU. This general dispersion may not be demonstrated where a plausible choice is drawing responses, because in that case it is quite possible for the lower-scoring group to find an attractive error choice even more seductive than is true for the more able group, thus resulting in a lower RU for the lower-scoring group. But a low RU attributed to a plausible mislead is not limited to lower-scoring candidates. For certain kinds of items, with very difficult antonyms as perhaps the prime example, the stem may be so unfamiliar to the lower scoring group that the potential seductive power of the most attractive error choice is lost on them, resulting in their answering in a more random-like way, and receiving a higher RU than somewhat more able examinees who knew the stem word but who could not educe the subtle relationship required to answer the item correctly. Distributional effects, then, are a useful but uncertain index of group behavior.

Mean RU's for items grouped by format and for examinees grouped by sex and ethnic group are given in Table 5. It is evident upon inspection that RU differences associated with sex are minimal. There are systematic differences by ethnic groups, however, with mean RU's tending to be least for Whites and greatest for Blacks. Pooling mean RU's across sex and item-format groupings, the ranges

Table 5

Mean Relative Uncertainty (RU) Values for Items within Item-Format
Computed from Error Responses of Original and Matched-Sample Examinee Groups

		Proportion of Examinee Groups*					
	Number of	Male			Female		
Item Format	Items	Wh	Ch	Bl	Wh	Ch	Bl
		Original Groups					
Verbal							
Analogies	18	81	89	89	81	84	89
Antonyms	20	87	89	93	85	88	93
Sent. Completion	17	77	86	89	75	83	89
Rdg. Comprehension	40	82	86	91	80	85	90
Quantitative							
Data Interpretation	14	69	78	79	71	75	79
Other	41	80	83	87	82	83	89
(No. of examinees)		(2000)	(195)	(915)	(2000)	(130)	(1715)
		Matched-Sample Groups					
Verbal							
Analogies	18		86	89		85	85
Antonyms	20		91	93		88	91
Sent. Completion	17		85	88		83	86
Rdg. Comprehension	40		87	89		86	88
Quantitative							
Data Interpretation	14		74	77		75	75
Other	41		83	85		83	85
(No. of examinees)			(570)	(350)		(390)	(360)

*decimals omitted

of mean RU's for the GRE-V are 75 to 87, 83 to 89, and 89 to 93 for Whites, Chicanos, and Blacks, respectively; for the GRE-Q, in the same order, they are 69 to 81, 75 to 83, and 79 to 89. In general, then, Blacks and Chicanos show more random responding than whites. These differences are again consistent with mean GRE-V and GRE-Q score differences, because there is a general tendency for lower scores to be associated with more random-like behavior, and thus higher RU values.

To help determine whether the mean RU differences are best explained on the basis of these mean score differences, or whether there are some systematic effects attributable, instead, to ethnic status itself, we may compare the 24 RU values for the matched sample to the 24 observed for the original minority groups. These turn out to be highly similar: of the 24 comparisons, 18 were only 02 or less, and the greatest difference observed was only 04. Just as for the Low P+ and the Reverse Omitting indices of guessing behavior, then, differences in RU apparently associated with ethnic group membership are more parsimoniously attributable to overall level of performance on the GRE-V and GRE-Q measures.

By far the sharpest differentiations among the mean RU's presented in Table 5 are those associated with item format. For all of the ten examinee groups, the highest mean RU's were observed for Antonyms (for Chicano males, Antonyms were tied with Analogies for highest mean RU). These mean RU's for Antonyms across the ten groups ranged from 85 to 93. Among the four GRE-V formats, the observed mean RU for Sentence Completion was as low or lower than any other item type for all but one group (white females matched to Blacks). Mean RU's for Sentence Completion items for the ten examinee groups ranged from 75 to 89. The lowest mean RU's of all were for the Data Interpretation item format, ranging from 69 to 79 across the ten examinee groups.

The most plausible interpretation for the higher RU's for Antonyms is that the more difficult Antonyms may involve stem words likely to be unknown to many examinees, leaving them little basis for answering systematically, whether correctly or incorrectly. Conversely, nearly all Sentence Completion items can be answered systematically (though not necessarily correctly) by using the immediate context to eliminate at least one of the less plausible choices. Thus, RU differences may reflect differences in the attractiveness of distractors, which the MnO index also reflected.

Consistent with the intermediate status of mean RU's for Analogies is the status of these items in relation to the bases they afford for making informed, systematic guesses. Although the vocabulary load in Analogies is considered secondary to the testing of sometimes subtle relationships, the stem words in some Analogies nevertheless appear to be extremely difficult. The two forces, vocabulary level and relationship difficulty, should tend to keep the RU's averaged across Analogies at some middle value between mean RU's observed for Antonyms and those for Sentence Completion items. Yet another factor likely to influence RU values for Analogy items is the presence of apparent differences in the strength of word associations between the stem and choice words of each item. Although word associations are not the appropriate bases for solving verbal analogies, there is evidence that they are often used, sometimes to the examinees' advantage (Moore, 1971; Willner, 1964), and sometimes to their disadvantage (Moore, 1971; Pike & Flaugher, 1970). The tendency to answer on the basis of word associations may be expected to lead to highly systematic error responses, and therefore to lower the RU's of analogies. However, there are two counteracting tendencies: 1) many examinee's errors are likely to reflect distractor effectiveness from the perspective of the quality of analogical relationship, rather than from word associations, and 2) the subset of Analogies having very difficult stem words would be expected to have only weak associational effects because these words are likely to be unfamiliar to many examinees.

As noted above, mean RU's for Data Interpretation items were markedly lower than for any other item type. Among the ten examinee groups, the median value of the mean RU's for these items was 75; that for the second-lowest set, "Other Quantitative," was 83, and those for the remaining four item formats ranged from 85 to 90. Two likely reasons for the lower mean RU's for Data Interpretation are the following: first, the Data Interpretation items require answering at two levels--the table or graph must first be interpreted in respect to a particular question, and then the computations or quantitative reasoning required by the more typical discrete items must be carried out. If there is a tendency for examinees to eliminate one or more of the choices at the first of these levels, and then to make a final choice based on the second, there would be a tendency for a concentration of the error choices among those not eliminated at the first level. This would, of course, yield lower RU values. Second, there is likely to be a subset of Quantitative items of both formats (Data Interpretation and "Other Quantitative") having only a few possible answers that are logically plausible. These items would then have one or more choices that would attract only the very least informed examinees. As a result, error choices would be concentrated among three or fewer distractors, which would lead to low RU's.

The demonstration of these differences in RU for mathematics items format should not obscure the consistency in the performance of ethnic groups when ability level is adjusted. That is, the item format patterns are demonstrated for all groups. Since item format patterns are predicated upon the group's approach to the item material, similarity in pattern is an indication of similarity in approach. The general impression through analysis by P+R, P+A, MnO, and RU is that the various ethnic groups are highly similar in approach, arguing for similarity in treatment in the instruction and scoring and for an evaluation of the present scoring and instruction as unbiased.

Findings for Selected Difficult Items

Although generalizations can be made relating item format to the indices of guessing behavior, there is considerable variation among individual items within the format groupings. A more detailed level of observation is needed to explore possible supplementary generalizations and clarifications based upon specific items. These item-level observations must be made within the framework of an exploratory analysis which takes advantage of a large body of data, but which is only indirect evidence for the inferences regarding those guessing strategies or tendencies that influenced the examinees' GRE responses. The results of this enquiry may be very helpful in designing and carrying out subsequent studies.

For selected individual items, data for each of the guessing-related indices (percent-pass (P+R and P+A), mean scores of omitting subgroups (MnO), and relative uncertainty (RU)) were examined. For most item formats, items were selected for this level of observation only if Low P+R or Reverse Omitting (indicated by MnO > 13.5) was evidenced for the given item by at least one of the six original examinee groups. The only exceptions were Sentence Completion and Reading Comprehension items, for which instances of Low P+R and Reverse Omitting were rare. For each of these item formats, the six most difficult items were selected for the more detailed examination, even if there was no Low P+R or Reverse Omitting for those items.

To facilitate presentation and discussion of the item-level guessing data, those have been grouped according to whether the items are verbal or quantitative, and whether the examinees are male or female.

Verbal items, male examinees: P+R. Data regarding the four guessing-related indices for individual GRE-V items, based on responses from males in the three original examinee groups (Whites, Chicanos, and Blacks), are given in Table 6. In comparing the P+R values in Table 6 across examinee groups, it should be recalled that mean corrected scores on the 95-item GRE-V for male Whites, Chicanos, and Blacks were about 49, 34, and 26, respectively, and that the P+R index as a measure of item difficulty would reflect these mean differences. That is, the expected pattern of P+R values for any given GRE-V item across the three groups would be one of diminishing P+R when moving from White to Chicano to Black, if the observations made for items grouped by format hold for individual items, as well.

There are signals of ethnic-group differences at the level of individual items, two of which are not linked to the general ability effects observed at the test level. First, there may be inversions in the rank-ordering of item difficulties across examinee group. Second, culture-bound aspects of some items could result in unexpectedly large differences in difficulty, such that P+R values for majority-group examinees of, say, 60, might drop well below the chance range for minority groups. An inspection of P+R values in Table 6 is reassuring on both points. Of the 28 items, 18 conform exactly to the P+R pattern of White > Chicano > Black. In the other cases, only one of the three relationships is reversed, with the largest reversal only 05. The underlying similarity of P+R values across the three grouos is underscored by the fact that in all but two instances of Low P+R for minority examinee groups, the P+R values for Whites were also 25 or lower.

A comparison between the P+R values observed for males in the original minority examinee groups and the matched groups drawn from the White male sample may be made by comparing the Chicano and Black data in Table 6 to those for the respective matched groups, presented in Table 7. Even at this item by item level of comparison, the correspondence between P+R values for the original Chicano male sample and the matched White male samples is very close, further confirming that observed White versus Chicano differences in P+R can be readily attributed to differences in total GRE-V score, rather than to factors specifically related to ethnic group membership, such as bilingualism. The P+R values across individual items for the Black group and the White group are also strongly comparable.

<u>Verbal items, male examinees</u>: <u>P+A</u>. Tables 6 and 7 also refer to P+A. The P+R and P+A values are conceptually and empirically related, in that the former is the percentage of examinees reaching an item who answer it correctly, and the latter is the percentage of examinees answering an item who did so correctly. For a given examinee group, the values for the two indices will usually be close, but they may be quite divergent, depending on the number of examinees omitting the item. To illustrate, compare White male

Table 6

Selected Guessing-Related Indices for Difficult GRE-V Items, for White, Chicano, and Black Males

	White (N = 2000)				Chicano (N = 195)				Black (N = 915)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
					Analogies (18 items)							
5	36	42	10.4	68	21	28	12.0	81	22	28	12.6	81
6	27	41	12.1	99	19	28	13.0	94	12	16	14.3	97
7	14	37	12.4	94	07	15	13.0	99	08	15	13.9	99
8	10	15	12.6	97	06	09	14.0	98	08	10	14.4	99
9	12	14	12.6	82	09	11	13.5	94	09	10	14.1	96
32	34	51	11.1	88	23	36	12.3	95	16	26	13.0	98
34	43	47	10.9	68	16	18	11.6	77	16	19	13.0	85
35	18	41	12.1	96	15	31	12.6	92	09	17	13.7	94
36	08	23	12.7	90	08	19	13.3	99	06	14	13.9	99
37	07	11	11.5	86	10	16	12.4	91	08	13	13.1	88
					Antonyms (20 items)							
16	36	54	11.7	92	25	43	13.2	98	26	38	13.6	99
17	26	54	12.0	90	24	46	12.6	94	20	35	13.6	96
18	25	53	12.6	98	23	43	13.8	98	17	29	14.0	97
19	09	19	13.1	89	09	15	14.0	93	09	14	14.6	92
44	24	56	12.1	94	15	38	12.6	95	10	24	13.3	96
45	22	55	12.2	99	14	39	13.0	96	11	28	13.4	98
46	23	57	11.8	92	14	40	12.9	91	11	28	13.2	97
47	15	38	12.4	81	13	32	12.7	87	12	29	13.5	96
					Sentence Completion (17 items)							
26	53	59	10.7	86	40	44	12.1	93	35	40	12.3	96
27	49	55	11.1	87	47	51	12.6	99	29	33	13.0	98
28	39	50	11.1	77	17	23	12.3	81	17	23	13.0	89
53	46	69	11.1	80	24	45	12.0	77	19	40	12.4	86
54	40	63	11.4	84	22	42	12.3	88	21	46	12.5	93
55	19	34	11.6	57	10	20	12.1	73	07	16	12.7	85
					Reading Comprehension (40 items)							
75	55	57	10.9	74	35	38	10.8	82	29	30	12.0	90
78	37	42	11.0	89	31	35	12.2	89	28	31	12.8	91
86	64	66	10.2	91	41	43	11.2	95	31	34	13.0	98
88	37	38	11.7	93	17	18	11.2	99	22	23	13.4	99
92	42	45	11.0	85	26	26	12.5	88	20	22	13.0	86
95	49	49	----	86	28	28	----	94	31	31	----	99

Note: Low P+ values, defined as P+ ≤ 16, are underlined, as are instances of Reverse Omitting, defined as mean-omit scores (MnO) ≥ 13.5.

Table 7

Selected Guessing-Related Indices for Difficult GRE-V Items, for White Males Matched to Chicano and Black Males

	Matched to Chicanos (N = 570)				Matched to Blacks (N = 350)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
			Analogies (18 items)					
5	23	29	12.0	77	20	25	13.2	85
6	15	23	13.8	98	15	21	14.4	92
7	06	16	13.5	98	05	11	14.0	94
8	07	11	14.0	99	07	09	14.5	99
9	10	13	13.2	89	07	08	13.2	93
32	21	36	12.3	94	21	34	13.8	96
34	24	27	12.5	80	15	17	14.2	87
35	06	17	13.2	97	04	09	14.1	96
36	07	18	13.4	95	06	13	14.1	98
37	07	13	12.8	82	08	12	14.2	87
			Antonyms (20 items)					
16	26	42	12.8	99	20	31	13.3	99
17	15	34	13.2	85	13	24	13.8	86
18	20	40	13.5	99	19	30	14.6	99
19	06	11	14.0	91	07	11	14.1	94
44	13	34	13.1	90	11	26	13.7	84
45	13	40	13.1	97	11	28	13.6	96
46	13	39	13.0	94	17	48	13.5	88
47	12	31	13.2	92	12	24	13.8	95
			Sentence Completion (17 items)					
26	38	45	11.7	93	37	42	12.7	96
27	31	35	12.4	98	27	30	13.6	98
28	21	29	12.6	84	18	24	13.2	89
53	24	45	12.4	84	17	33	12.8	92
54	25	49	12.4	92	15	33	12.6	91
55	08	17	12.7	77	10	24	12.9	77
			Reading Comprehension (40 items)					
75	37	39	12.4	81	32	35	12.9	84
78	25	30	12.6	92	22	26	13.4	90
86	50	52	10.6	96	34	37	11.1	99
88	22	24	12.1	99	18	20	11.6	99
92	26	28	11.4	89	18	20	12.5	91
95	33	33	----	95	25	25	----	98

responses to GRE-V items 9 and 36, in Table 6. For item 9, the P+R and P+A values were 12 and 14, the slight difference due to a 15 percent omitting rate; for item 36, a much larger omitting rate of 65 percent resulted in a change from a P+R of 08 to a P+A of 23. Because of the differences between P+R and P+A resulting from omitting, a correspondence between P+A and mean GRE-V score like that between P+R and GRE-V score, does not necessarily hold. Nevertheless, when the P+A values in Table 6 are observed across examinee groups, item by item, the same basic pattern of values (Whites > Chicanos > Blacks) is in fact demonstrated. This indicates another level of consistency of behavior for which mean scores, regardless of sex or ethnicity, provide the key. However, it also indicates a kind of logical inconsistency, in that the rate of omitting is generally comparable across groups despite the differing amounts of information associated with examinee group.

Contrasts in P+R and P+A values for individual items may next be considered, particularly as these contrasts are related to item format. It was noted earlier that the proportions of Low P+R items that were also Low P+A items differed noticeably depending on item format; very difficult (Low P+R) Analogies were more often also items with worse-than-chance answering (Low P+A) than was the case for Antonyms. In examining specific P+R values, rather than distributions of Low P+R's, it is evident that part of the phenomenon relating to Analogies was a result of the typically lower values of P+R occurring for this kind of item. P+A is usually greater than P+R. If an item type has a number of very low P+R items, these are more likely to show values of P+A ≤ 16. In examining the P+R values across the three male subject groups in Table 6, for example, there are nine instances of P+R values of 08 or less involving Analogies, but only one for all other item formats combined. For these nine items, of course, omitting rates of at least 50 percent were required before the Low P+R values were raised beyond the threshold for Low P+A of 16. Average level of P+R, then is a clear factor in determining the frequency of Low P+A.

However, a detailed examination of (P+R, P+A) differences in Table 6, and a consideration of the omitting underlying these differences gives clear evidence that there are marked variations in the amount of guessing (as indicated by non-omitting) associated with a given range of item difficulty, depending on the item-format involved. These differences in guessing or risk-taking tendency associated with item format, then, also account for the relatively greater amount of worse-than-chance guessing that was observed for Analogies than for Antonyms.

The differences between P+R to P+A, and the omitting rates underlying these may be summarized for the four GRE-V item formats, based upon Table 6. By far the least omitting, and presumably the most guessing, was exhibited for items in the Reading Comprehension format. Among the six most difficult Reading Comprehension items, the omitting rates ranged from four to fourteen percent, across the three original male examinee groups, remaining low even for the most difficult item (88), which had minority P+R values of 17 and 22. These low omitting rates were necessarily associated with very modest P+R to P+A difference, as is evident in Table 6. Conversely, the most omitting, and thus presumably the least guessing, was observed for difficult items in the Antonyms format. Associated with the 24 between-group differences for Antonyms shown in Table 6, the omitting rates ranged from 29 to 65 percent, with a median of 54. Thus, typical P+A values for Antonyms were about twice the size of their P+R counterparts.

For the six most difficult Sentence Completion items, the picture is mixed. For the first three items, low omitting rates ranging from eight to 24 percent were observed, whereas omitting for the last three items ranged from 33 to 57 percent. Again, P+R to P+A differences were commensurate with the relative amount of omitting observed. It was suggested earlier that the Sentence Completion format, with its immediate contextual clues, would be expected to encourage guessing and therefore to reduce omitting. The most likely explanation for the high omitting rates for the last three Sentence Completion items is that they are also the last items of a separately-timed section of the GRE-V, and that the higher omitting rates are reflecting a speededness component of the test.

Omitting rates among the eight most difficult Analogies also varied substantially across items. For items 5, 9, and 34, P+R and P+A differences indicated in Table 6 for the three male groups were modest, deriving from percentages of omitting ranging from 8 to 26. For items 7, 35, and 36, on the other hand, large differences are evident, associated with percentages omitting which ranged from 49 to 65. The other two Analogies were intermediate in the P+R - P+A differences and in rate of omitting. The most likely basis for this wide variation in P+R - P+A differences and omitting rates among the difficult Analogies is the one cited earlier in a discussion of the mean RU's observed for Analogies: the considerable range of apparent difficulties of words used in the stems of these items. Even very difficult Analogies are attempted by a large proportion of the examinees, provided that the vocabulary load for the items is not too great. For items having very difficult stem words, however, examinees seem much less likely to venture a guess, and thus much more likely to omit the item.

Verbal items, male examinees: MnO. In this study, when those omitting a given item had a mean GRE-V score at the group's fifty-fifth percentile, or higher, the phenomenon of Reverse Omitting was considered to be present. Since mean scores of examinees omitting a given item were scaled separately for each group, with the overall group mean for a given group set at 13.0 and the standard deviation at 4.0, the MnO scale is then related to the percentile scale as illustrated by the following (MnO, percentile) pairs: 12.0, 40; 12.5, 45; 13.0, 50; 13.5, 55; and 14.0, 60. Reverse Omitting is thus defined as MnO $\geq$ 13.5.

Reverse Omitting behavior at the item level, across the male White, Chicano, and Black examinee groups, is reflected in Table 6. There are fifteen instances of Reverse Omitting in this table, indicated by underlined MnO values. The correspondence of occurrences of Reverse Omitting to those of Low P+ values is evident. Nine of the Reverse Omitting instances are associated with P+R values of 10 or less; the highest associated P+R value was 26, just above chance level. The evaluation of MnO, the basic phenomenon in Reverse Omitting, is exhibited most frequently in difficult items.

This correspondence between Low P+ and Reverse Omitting, however, holds more for minority groups than for Whites. To make this point, a somewhat relaxed criterion of reverse omitting may be used, MnO $\geq$ 13.0. Only one of seven instances of Low P+R for Whites was accompanied by MnO $\geq$ 13.0; for Chicanos, six of 13 instances were so accompanied, and for Blacks, 14 of 15. This progression could be due to chance differences, however unlikely. They could, alternatively, be due to differences in the omitting strategies of examinees confronted with very difficult test items, differences characteristic of the different ethnic backgrounds. A third possibility is that the different ratios of high MnO to Low P+ occurrences may be in some way a function of the increasing relative difficulty for a given item across the three groups, an increase not apparent in either of the P+ values provided.

There is sufficient complexity in this discussion to warrant a restatement. The most direct data for success on an individual item is the P+R column in Table 6. The data show that an item such as #36 is succeeded on at approximately the same basic rate by each ethnic group. But the MnO values for the groups are different. Whites who omit are not more able than the average White. Blacks and Chicanos who omit are above the average for their group. These differences may be attributable to chance, to ethnic group behaviors, or to some complex process reflecting group differences in ability. That is, while the groups are equal in "ability" on item 36, they are known to differ in ability on the test. Do these test-score differences account for the MnO differences?

The most direct answer to this question is derived from the matched sample data. Comparing the MnO data for the minority examinees in Table 6 to those for matched-sample examinees in Table 7, shows that observed MnO values, including those in the Reverse Omitting range, correspond closely, item by item, for the original Chicano group and its matched White counterpart. The MnO values observed for the White males matched to the Black male examinees generally tended to be moderately higher, item by item, than those observed for the original Black male group, but the comparability is strongly established, and in fact the incidence of Reverse Omitting increased from seven items, for the Black group, to 16 items, for its matched White counterpart. The apparent ethnic difference in omitting is an artifact of differences in group ability. While Table 7 establishes that the rising MnO values are linked to ability level, an attempt at our explanation of the phenomenon is appropriate. In most instances, an item that has comparable and very difficult values of P+R for two given groups will in some real sense be even more difficult for a group having a lower mean score on the overall test. It could be that for a given set of test items a greater amount of Reverse Omitting for a lower scoring group is due to this increase in "true" item difficulty. The following section attempts to demonstrate this.

Consider first some theoretical relationship between ability and MnO level. If for a given group a test item is very easy, only the lowest scoring examinees in the group would need to consider omitting the item, and a MnO value of, say, 9.0, might reasonably be expected. For a less able group the need to decide whether to omit the item might be distributed over more of the examinees, such that a value of MnO of 12.0 would be expected. MnO values for a given test item should increase as a function of decreasing group score means. They should not, in general, exceed 13.0, but should approach it as a limit when the item is too difficult for essentially all examinees in a given group. By this model, P+R and MnO are associated; as P+R drops, MnO rises. But this model is an items-level model. Suppose in addition that omitting behavior is governed by test-level constraints.

The total amount of omitting that examinees do appears, in fact, to be unrelated to their total test scores (Swineford & Miller, 1953; Slakter, Crehan, & Koehler, 1975). This would suggest that a group's tendency to omit may be relatively stable over some range of test difficulties limited to no more than, say, ten percent of the items in a test. The omitting behavior of the highest-scoring and lowest-scoring fifths of an examinee group, then, with respect to the most difficult fraction of the items is likely to be somewhat different. On most tests, omitting for the most able group should be concentrated among these most difficult items. However, many more of the items in the same test are in the "omittable" range for the examinees in the lowest fifth of the group. In that case, if examinees in this subgroup still tend to limit omitting to about ten percent of the total number of items, it is likely that their omitting will be widely dispersed across a larger group of the items, rather than concentrated in the hardest few. The net result would be, for each of the hardest items, a greater relative omitting rate on such an item on the part of the higher scoring examinees than would be observed for the lower scoring ones. This could in turn yield MnO values in the Reverse Omitting range for these items, for the full examinee group.

In general, then, the apparent differences in omitting behavior for the various groups, indicated by the rising values of MnO, is an artifact of ability level distinctions, confirmed as such by the data in Table 7 for matched groups and supported by a rationale which sees low-level examinees omitting fewer items than would be expected on the basis of the general relationship between item difficulty and omission.

<u>Verbal items, male examinees</u>: <u>RU</u>. The additional information derived at this level of observation for the RU index offers little or no interesting patterns. Thus, the RU data will not be discussed in greater detail than was done earlier, in the form of mean RU values for data averaged across items grouped by format, within examinee subgroup.

<u>Verbal items, female examinees</u>: <u>P+R</u>. Data regarding the four guessing-related indices for individual GRE-V items, based on responses from females in the three original White, Chicano, and Black examinee groups, are given in Table 8. When applying these indices to the data obtained from the female examinees, the data and their interpretations already reported for males will be taken as the point of departure. To the extent that similar observations hold for the female examinees, we have essentially a replication of the male examinee findings.

When comparing the P+R values in Table 8 across examinee groups, it is useful to recall that the mean corrected scores on the 95-item GRE-V for female Whites, Chicanos, and Blacks were about 49, 33, and 24, respectively. These values depart only slightly from those for the male examinee groups. Thus, the expected pattern of P+R values for any given item would be highest for White examinees, intermediate for Chicanos, and lowest for Blacks, unless there were item-by-ethnic group interactions such that for some items these

Table 8

Selected Guessing-Related Indices for Difficult GRE-V Items, for White, Chicano, and Black Females

	White (N = 2000)				Chicano (N = 130)				Black (N = 1715)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
Analogies (18 items)												
5	24	30	11.0	70	15	21	10.9	80	15	21	13.1	86
6	44	55	11.4	99	27	35	12.9	97	17	21	13.9	94
7	20	46	12.1	92	11	21	12.5	96	09	17	13.7	99
8	11	16	12.4	98	10	13	12.9	99	06	08	14.3	99
9	13	16	12.8	82	09	14	12.0	96	12	14	14.1	97
32	32	48	10.9	87	17	29	12.0	98	14	23	13.3	98
34	49	53	11.0	72	26	29	10.7	74	16	18	12.7	87
35	21	47	11.9	92	19	36	11.8	77	07	15	13.8	93
36	10	24	12.3	85	06	16	12.7	90	07	13	13.9	98
37	11	16	11.2	83	07	11	12.2	84	08	12	13.2	89
Antonyms (20 items)												
16	48	61	11.0	91	35	56	12.0	90	28	40	13.3	97
17	25	52	12.1	85	19	36	12.4	95	16	30	13.7	95
18	22	45	12.5	97	25	48	12.5	95	15	26	13.8	97
19	11	20	13.1	83	12	20	12.8	92	08	12	14.3	92
44	32	63	11.5	95	25	47	12.3	81	12	25	13.3	96
45	19	54	12.2	99	15	39	12.3	98	11	29	13.5	94
46	27	60	11.5	83	18	51	12.3	89	11	27	13.2	97
47	17	41	12.3	84	16	35	12.5	78	12	26	13.4	97
Sentence Completion (17 items)												
26	50	54	10.3	84	28	34	11.4	87	34	38	12.3	98
27	54	59	10.9	85	39	44	9.9	99	30	34	13.3	99
28	42	53	11.5	81	23	33	11.7	87	17	23	13.1	89
53	47	69	11.0	81	28	53	11.7	90	21	38	12.7	88
54	38	60	11.2	82	22	49	12.0	79	22	42	12.8	97
55	18	29	11.3	47	13	33	12.0	70	08	16	12.9	81
Reading Comprehension (40 items)												
75	59	61	9.8	74	47	49	8.6	83	27	28	12.8	88
78	38	42	11.1	86	42	46	10.6	85	27	31	12.6	90
86	57	60	11.8	90	40	45	11.3	90	28	30	12.6	99
88	32	34	11.8	96	24	26	12.1	98	17	19	13.3	99
92	48	50	11.8	77	28	29	16.0	82	21	23	12.4	87
95	45	45	----	84	37	37	----	94	22	22	----	98

patterns were reversed. As noted earlier, there were only a few reversals for the males. For the female examinees, there were only seven reversals among the 84 comparisons, none of them exceeding P+R differences of 06. The other possible ethnic-group effect in P+R values to be considered is that of an unusual difference in P+R between the majority-group examinees and either of the minority groups, with the latter in the Low P+R range of 16 or less. For the males, as noted earlier, there were no such differences; all instances of Low P+R for minorities were accompanied by P+R values of 25 or less for the Whites. Differences were somewhat more pronounced for the female examinees, but still not to a degree suggesting group-item interactions. For the 17 GRE-V items having a low P+R value for at least one of the minority female groups, the median P+R for majority-group females was 20. For 16 of these 17 items, the White P+R was 32 or less; the exception was a P+R value of 49. Even in the latter two instances, intermediate values on the part of Chicano females suggested an underlying similarity such that the differences appeared more likely to be associated with total-score differences among the three ethnic groups than with categorical differences of some sort associated with ethnic-group membership per se. Thus, the P+R values for female Whites, Chicanos, and Blacks were 44, 27, and 14 for item six, and 49, 26, and 14 for item 34.

A comparison between the P+R values observed for females in the original minority examinee groups and the matched groups drawn from the White female sample may be made by comparing the Chicano and Black data in Table 8 to those for the respective matched groups presented in Table 9. As was true for male examinees, the correspondence at this item-by-item level between P+R values for the original minority-group examinees and the groups matched on GRE aptitude scores is generally close, further indicating that observed differences between the majority and minority-group examinees can be attributed to total GRE-V score differences more readily than to factors related to ethnic group differences such as bilingualism.

Verbal items, female examinees: *P+A*. P+A values are related to their P+R counterparts according to the percentage of examinees omitting each item. Thus, the P+R and P+A values for White female responses to item 9 are quite close, at 13 and 16, while those for item 36 are quite different, at 10 and 24, the contrast being attributable to an omitting rate of only 18 percent for item 9, but of 57 percent for item 36. Because of such differences in omitting, P+A patterns across examinee groups could well differ from P+R patterns, particularly if there were differential guessing tendencies associated with examinee group. When the P+A values are examined across examinee groups, however, the same basic pattern observed for male P+R values (Whites > Chicanos > Blacks) prevails.

Table 9

Selected Guessing-Related Indices for Difficult GRE-V Items, for White Females Matched to Chicano and Black Females

	Matched to Chicanos (N = 390)				Matched to Blacks (N = 360)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
			Analogies (18 items)					
5	11	16	12.0	72	10	14	12.4	81
6	27	37	13.0	98	22	28	13.8	94
7	08	20	12.9	98	07	17	13.4	98
8	06	09	13.2	99	04	05	13.9	99
9	11	13	13.9	95	09	10	14.2	94
32	18	32	12.3	92	11	23	12.5	96
34	32	35	11.6	78	22	24	11.6	77
35	09	25	12.7	90	09	21	12.9	92
36	06	16	12.8	97	06	15	13.4	96
37	07	12	12.4	86	06	10	12.9	82
			Antonyms (20 items)					
16	29	40	12.4	90	28	39	13.0	96
17	14	31	13.2	85	14	29	13.0	88
18	18	34	13.2	95	14	28	13.8	98
19	06	10	13.7	86	08	13	13.8	90
44	15	39	12.7	88	11	26	12.7	92
45	13	41	13.0	99	09	30	13.0	94
46	09	29	12.6	92	09	26	12.7	93
47	11	29	13.1	91	13	32	13.1	88
			Sentence Completion (17 items)					
26	34	39	12.0	91	27	33	11.5	99
27	41	49	12.4	99	32	37	13.2	99
28	23	31	12.2	86	20	28	12.0	90
53	33	58	11.7	82	19	37	12.2	92
54	21	41	11.9	90	17	33	12.3	87
55	11	23	11.8	61	07	14	11.9	75
			Reading Comprehension (40 items)					
75	41	43	12.6	81	29	31	13.0	85
78	33	36	12.0	89	28	33	12.1	93
86	43	46	12.5	98	37	41	12.7	99
88	21	22	13.2	99	14	15	13.8	99
92	30	32	11.8	85	22	23	12.6	80
95	27	27	----	94	15	15	----	93

The differences between P+R and P+A, and the omitting rates underlying these differences may be examined for each of the four GRE-V item formats. Table 8 provides the initial observations for females, paralleling the data given in Table 6 for males. As was true for data in Table 6, by far the least omitting, and thus presumably the most guessing is shown by Reading Comprehension items. Omitting for the six most difficult Reading Comprehension items, across the three original female groups, ranged from three to 14 percent. Consistent with these low omitting rates are the consistently modest P+R to P+A differences. Conversely, the most omitting and thus the least guessing was again observed for Antonyms, with omitting for the eight Antonyms listed in Table 8 ranging from 22 to 68 percent.

Sentence Completion and Analogies data in Table 8 were also very similar to those in Table 6. For the first three Sentence Completion items, omitting rates ranged from seven to 30, in contrast to the last three items, with omitting rates ranging from 32 to 60 percent. As noted for the male responses, the low omitting rates are consistent with the observation that the immediate contextual clues provided by the Sentence Completion format would be expected to foster guessing and thus to reduce omitting. The high omitting for the last three items is again likely to be a function of their being the last three items in a separately timed test section. Omitting rates for Analogy items 5, 9, and 34 were again relatively modest, from 14 to 34, which may be contrasted to those for items 7, 35, and 36, which ranged from 46 to 60. Since the same items tend to contrast in omitting rates for females as for males, the same rationale again seems most plausible; the difficulty of the stem words seems to have a direct bearing on the likelihood that examinees will omit rather than attempt a difficult Analogy item.

Verbal items, female examinees: MnO. The eleven instances of Reverse Omitting (MnO > 13.5) by female examinees are underlined in the MnO columns of Table 8. The correspondence of Reverse Omitting to low P+R values is agaiu evident; P+R values associated with instances of Reverse Omitting ranged from 06 to 16.

To examine the correspondence between Low P+R and Reverse Omitting differentially for the three female examinee groups, it is again helpful to use a more relaxed criterion of reverse omitting. Using a criterion of MnO > 13.0, we observe that for Whites, one of five instances of Low P+R was accompanied by a value of MnO above this level, compared to none of 10 for Chicanos, but 15 of 17 for Blacks. These patterns (1:5, 0:10 and 15:17) across ethnic groups among female examinees may be compared to that noted earlier for male examinees, for whom the ratios were 1:7, 6:13, and 14:15 for the same sequence of ethnic groups. Once again, there is an apparent difference by ethnic group in the correspondence between Low P+R values and patterns of omitting behavior. As the discussion of the data for males showed, these differences coul: be due either to differences in omitting strategy of examinees confronted with very difficult items that are characteristic of the separate ethnic groups of female examinees, or they could be in some way a function of relative difficulty of the test as a whole for the three groups, with constraints on total omitting leading to decreases that are not apparent in either the P+R or the P+A values provided. In the White-Chicano comparisons, however, the tendency for MnO differences to follow total score differences is much weaker than expected. Mean GRE-V scores for White, Chicano, and Black females were 49, 33, and 24, respectively. Yet, among the 30 White-Chicano comparisons of MnO values, there were nine reversals, cases where White MnO index exceeds Chicano values, and an inspection of all 30 pairs of MnO values indicates a general and inordinate similarity in these values for the two examinee groups. The reversals are quite uniformly distributed across the four-item formats, so these do not appear to be a factor. A re-examination of White-Chicano comparisons of MnO values for males shows that while the GRE-V means for three male groups, 49, 34, and 26, closely approximate those for the female groups, and the White-Chicano, there are only two reversals, and differences in MnC are consistent, both in direction and magnitude, with expectations. The White-Chicano results in MnO comparisons for females, then, are different. While MnO values for White males are very similar to those for White females, and those for Black males to those for Black females, among the Chicano examinees, females consistently tended to be one or two points (one-fourth to one-half standard deviations) lower than males. Given the very similar GRE-V means for Chicano males and females, this is not likely to be the cause. The slightly larger standard deviations in GRE-V scores for the Chicano females than for Chicano males might account for some of the MnO difference, but probably for very little. A greater tendency for female Chicanos than males to omit the more difficult items would readily explain the MnO differences, but an inspection of percent-omitting on each of the items for which the MnO reversals were observed showed very similar omitting rates, item by item, for the two examinee groups. Experimental data are needed to better establish and to adequately explain why the MnO values for female Chicanos were lower than expected.

In comparing the Table 8 and Table 9 MnO values, it is evident that the MnO values for the female Chicano group were not replicated for the White female group matched in total score distribution to the female Chicano group. The matched sample shows the expected increased values of MnO. Thus, the finding appears to be related to a characteristic of the Chicano female group not explainable on some other basis such as total scores on the GRE-V. This conclusion is reinforced when the original and matched-sample data for Chicano males, provided in Tables 6 and 7, are included in the comparison. There are strong similarities, item by item, in MnO values among the three data-sets, original and matched Chicano male, and matched Chicano female. Only the original Chicano female MnO values fail to fit the pattern. As noted earlier, they tend to be about one-fourth to one-half standard deviation lower than would be expected. In fact, summarizing MnO data across all 10 subject groups, the original Chicano female data are the only ones not fitting a well-defined and explained pattern. The cluster made up of White male and White female MnO values are strongly consistent, as are the data within the original and matched Black male groups and original and matched Black female groups. Further, there is a systematic between-cluster consistency, in that the between-cluster differences in MnO values are consistent with an explanation deriving from the different GRE-V score means associated with the three clusters (the third being the Chicano-related ones exclusive of the original Chicano female group). In summary, then, the MnO data for the GRE-V takes the form of a highly consistent data base for nine groups, with the exceptional group, Chicano female, needing further study.

Verbal items, female examinees: RU. As with the male sample, the findings at the level of the individual items add nothing to the earlier discussion of item format results.

Quantitative items, male examinees: P+R. Data for the four guessing-related indices for individual GRE-Q items, based on the responses of White, Chicano, and Black males, are given in Table 10.

Comparisons of the P+R values in Table 10 across the three examinee groups should consider that mean corrected scores on the 55-item GRE-Q for male Whites, Chicanos, and Blacks were about 33, 21, and 14, respectively. (Standard deviations were about 10.) The expected pattern of P+R values for any given GRE-Q item would be one of diminishing P+R when moving from White to Chicano to Black. This should generally hold, if between-group differences in item difficulty are essentially attributable to the general ability reflected in the total GRE-Q score. As noted above, however, between-item differences in P+R related to ethnic group differences may be such that the rank-ordering of item difficulties varies from one ethnic group to another. Second, culture-bound attributes of some items could result in unusual differences in item difficulty such that a high P+R value for the majority group may drop into the chance range for one or both minority groups.

Table 10

Selected Guessing-Related Indices for Difficult GRE-Q Items, for White, Chicano, and Black Males

	White (N = 2000)				Chicano (N = 195)				Black (N = 915)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
Data Interpretation (14 items)												
7	41	41	8.5	03	18	18	---	18	17	17	14.8	09
10	41	42	10.4	64	17	17	11.8	72	19	20	13.0	59
19	28	31	10.8	81	14	17	12.0	88	10	11	12.8	87
20	21	22	8.8	66	10	10	10.4	72	16	18	11.4	82
21	60	75	10.9	47	35	46	11.8	71	28	40	13.2	68
32	44	56	10.7	83	20	28	12.1	83	13	20	12.8	89
33	34	47	11.6	94	17	25	12.7	92	11	17	13.3	87
34	66	85	10.8	83	60	84	11.6	89	39	58	13.0	91
Algebra (10 items)												
27	65	67	9.8	74	44	46	10.0	80	34	36	12.8	87
39	58	63	10.7	68	28	32	11.3	84	22	25	12.2	85
41	73	86	10.9	87	46	58	13.3	84	40	46	13.7	71
53	41	56	12.6	98	09	13	14.6	94	17	22	14.5	99
55	14	14	----	72	01	01	----	84	06	06	----	90
Geometry (14 items)												
42	83	92	9.7	95	63	74	12.2	88	51	62	12.5	92
46	63	73	11.4	82	46	54	11.7	80	41	49	12.8	90
48	48	53	11.6	39	29	33	12.5	49	26	31	12.4	55
52	49	78	11.9	95	30	51	12.3	91	23	40	12.9	97
54	44	49	12.2	94	26	31	14.5	98	14	16	14.1	92
Arithmetic (12 items)												
45	49	63	11.4	82	28	41	12.7	94	26	34	13.5	95
49	46	49	11.2	47	29	31	15.3	54	33	36	12.1	79

Note: Five items were classified "miscellaneous," bringing the total number of quantitative items to 55.

When the P+R values in Table 10 are compared for the consistency of P+R differences between groups, the result is much the same as for the GRE-V data. Only five of the 60 between-group comparisons for the 20 items examined showed differences in the direction not expectrd, the largest of these P+R differences was 12. On the question of discontinuities in item difficulty associated with majority or minority ethnic status, the GRE-Q data depart from those for the GRE-V. It will be recalled that for male examinees, all but two instances of Lcw P+R for minority examinee groups on GRE-V items were accompanied by P+R values for Whites of 25 or lower. On the GRE-Q, however, five of seven items with Low P+R for one or both minority examinee groups were associated with P+R's for White males ranging from 48 to 44. These differences are related to test score for most of these items for the values of P+R are Chicano males about midway between those for Whites and Blacks. What remains clearly evident is that there is a much steeper gradient in P+R across examinee groups on difficult Quantitative items than is true of difficult Verbal items.

A comparison between the P+R values observed for minority male examinee groups and matching groups drawn from the full White male sample may be made by comparing the Chicano and Black data in Table 10 to those for the respective matched groups in Table 11. In making the comparisons, it should be noted that the mean scores for the two matched groups are each about three raw score points lower than those for the original minority group samples. This should mean an average drop between P+R and P+A of about 06, when going from a given minority group to its matched White counterpart. With that adjustment, P+R values for the White males matched to Chicanos are very close to the expected values for most of the items. The exceptions are generally within 10 of the expected P+R; upon examination of the particular items involved, there are no common characteristics that might suggest an explanation.

Quantitative items, male examinees: P+A. The P+A values will be identical to P+R if there was nc omitting, but will be quite divergent if there was a high percentage of omitting. To illustrate, compare White male responses to GRE-Q Geometry items 48 and 52, in Table 10. For item 48, P+R and P+A are 48 and 53, the small difference due to an omitting rate of 10 percent; for item 52, a larger omitting rate of 37 percent resulted in a change from a P+R of 49 to a P+A of 78. Because of the differences between P+R and P+A that can occur through omitting, the correspondence between P+R and examinee groups mean GRE-Q does not necessarily imply a similar correspondence between P+A values and the GRE-Q means. However, when the P+A values in Table 10 are observed across examinee groups, the same general pattern of values (Whites > Chicanos > Blacks) is observed for P+A as for P+R values and for GRE-Q means. Further, the three between-group comparisons in which P+A differences did not fit the general pattern were for the same GRE-Q items in which P+R differences were reversed. Thus, the pattern of item difficulties in the form of P+R values which closely reflected GRE-Q mean score differences is closely paralleled by the pattern of P+A values, a finding of significance because the latter are influenced directly by any systematic guessing tendencies.

Differences between P+R and P+A for the GRE-Q items may be associated with item format, as was true for GRE-V items. Summarizing over the three original and two matched male examinee groups for Data Interpretation items, there were 13 instances of Low P+R items, 7 of which were also Low P+A. The relative numbers of Low P+R and Low P+A items was about the same for Other Quantitative items. Of 9 Low P+R items, eight were also Low P+A. The consistency of Low P+R and Low P+A comparisons for both kinds of GRE-Q items was due both to very low P+R values (the median was only 12), and to generally low omitting rates despite the high level of difficulty of the items (the median percent of examinees in each group omitting these items was about 20).

Quantitative items, male examinees: MnO. The next index of guessing behavior in the responses of male examinees to the GRE-Q is that of Reverse Omitting, indicated by group mean-omitting (MnO) values of 13.5 or greater. There are eight such instances of Reverse Omitting indicated in Table 10 by underlined MnO values. A correspondence of occurrences of Reverse Omitting and of Low P+R values was found with GRE-V items for both males and females but was not as strongly evident in male responses to the GRE-Q. The maximum P+R Reverse Omittings on GRE-V items by males was 26, and most P+R values were less than 10; all eleven Reverse Omits on the part of females were associated with P+R values of 16 or less. In contrast, for the eight Reverse Omitting instances in Table 10, associated P+R values were 09; 14, 1717, 2626, 29, and 40. Four of these six values are inordinately high by the Verbal results.

The relation between Reverse Omitting and item difficulty may be quantified by considering the proportion of Low P+R's for which Reverse Omitting occurs. To provide a more stable number of instances of Reverse Omitting to work with, the criterion may be relaxed, as in earlier sections of this report, to MnO > 13.0. Between-group differences were noted in the Reverse Omit to Low P+R ratio for GRE-V item, with observed ratios for White, Chicano, and Black males of 1/7, 6/13, and 14/15, and in the same order, for females of 1/5, 0/10, and 15/17. A similar, but less pronounced pattern occurred for the same groups of male on the GRE-Q; 0/1, 1/4, and 5/7. As in earlier parts of this report, these differences are attributed to the increasing relative difficulty of any given item across the three groups that may not be apparent in the P+R values. The rationale for this attrition was given in the discussion of GRE-V data from male examinees, where it was also noted that that explanation would have an increased credibility if there were a general tendency, across a wide range of MnO values, for MnO to increase as examinee group mean scores decrease.

The MnO values in Table 10 were in the expected direction, White > Chicano > Black, for 56 of the 60 comparisons. This relationship supports the explanation that between-group differences in MnO are attributable to a general effect in which lower-scoring examinees are more likely to find items difficult but to exhibit higher MnO's because their omitting is inhibited. As suggested in the discussion for male responses to the GRE-V, it cannot, in itself, rule out the counter-explanation that the observed differences in Reverse Omitting are in some way associated with systematic differences in omitting behavior related to ethnic group (such as differences in risk-taking, for example). The MnO data provided in Table 11, however, for White males matched to the Chicano and Black male groups, provide a test of these alternatives. If the differences in Reverse Omitting are primarily a function of ethnic group membership, there should be a very low rate of Reverse Omitting for the matched samples drawn from the White group. On the other hand, if the effect is that of low-scoring groups encountering a large enough proportion of difficult items to yield a Reverse Omitting effect, then the incidence of Reverse Omitting should correspond to differences in the GRE-Q means of the five male groups involved.

In comparing the MnO data in tables 10 and 11, the explanation of Reverse Omitting in terms of numbers of difficult items is the one supported. The GRE-Q mean scores were, in descending order of magnitude: Whites, 33; Chicano-matched Whites, 24; Chicanos, 21, Black-matched Whites, 17; and Blacks, 14. With groups listed in the same order, the difficult-item hypothesis would predict Reverse Omitting frequencies going from low to high. Those observed were: 0, 1, 3, 2, and 5. To have more instances for comparison, the more relaxed criterion of MnO > 13.0 may again be used. Then, the frequencies of reverse omitting were: 0, 2, 4, 5, and 7, again consistent with the low-to-high hypothesis.

Quantitative items, female examinees: P+R. Guessing related data for individual GRE-Q items, based on the responses of White,

Table 11

Selected Guessing-Related Indices for Difficult GRE-Q Items, for White Males Matched to Chicano and Black Males

	Matched to Chicanos (N = 570)				Matched to Blacks (N = 350)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
Data Interpretation (14 items)								
7	23	23	---	07	13	13	---	13
10	24	25	10.2	61	20	21	8.3	66
19	14	16	10.9	84	10	12	11.3	86
20	12	13	10.2	73	07	08	10.0	80
21	35	51	11.8	63	27	40	12.0	73
32	23	32	12.0	83	14	20	12.9	87
33	18	27	12.6	88	14	19	13.0	86
34	55	76	12.0	88	46	63	12.2	92
Algebra (10 items)								
27	46	48	10.9	78	34	35	9.8	78
39	35	38	11.7	76	24	29	11.6	88
41	54	65	12.9	81	38	45	13.3	73
53	24	34	13.7	98	15	19	14.2	95
55	06	06	----	86	04	04	----	96
Geometry (14 items)								
42	73	84	11.4	95	59	71	11.5	98
46	48	55	12.1	88	38	44	13.1	96
48	35	39	12.0	46	25	28	12.8	54
52	34	56	13.0	93	27	45	13.2	92
54	27	30	13.3	94	15	16	14.9	94
Arithmetic (12 items)								
45	37	50	12.4	91	21	29	13.0	90
49	32	35	12.3	56	23	25	12.7	72

Chicano, and Black female examinees are given in Table 12. Comparisons of the P+R values across the three examinee groups may be made, keeping in mind the mean corrected scores on the GRE-Q for female Whites, Chicanos, and Blacks of about 26, 18, and 14, (standard deviations about nine). The expected pattern of P+R values for individual items is the same as that observed for GRE-Q means if between-group differences in item difficulty are primarily attributable to the general ability reflected in the GRE-Q scores.

An examination of P+R values in Table 12 indicates only five between-group differences in P+R in the reverse direction of the GRE-Q mean scores, out of sixty such comparisons. Most of the reversals occurred in comparisons between Chicano and Black P+R values, and in all but one instance, both values being compared were at or below the chance level of 20. In the chance region, of course, reversals can readily occur, since some items give meaningful P+R scores of five or even less, but 20 is the expected value if a group is answering in a purely chance manner. The steep gradient across the three ethnic groups noted earlier for male responses to GRE-Q items was not evident for the female examinees, primarily because White females showed consistently lower P+R values than their male counterparts.

The P+R values for minority female groups shown in Table 12 may be compared to those for White female groups selected to match them, given in Table 13. The comparison is facilitated by the fact that the GRE-Q scores for the White matching groups are nearly the same as those for the minority groups. Item by item, the match in P+R values between the Chicano and the Chicano-matched White group was very close for all but one item and there the difference was only eight points. The comparisons between original and matched Black female groups were nearly as close; there were modest differences for three of the 20 items, of eight or nine points.

Quantitative items, female examinees: P+A. While there were substantial differences between items in percent omitting, allowed for sizeable differences between I+R and P+A patterns for some items the percent of examinees omitting a given item varied little from one group to another. For example, on item 10, the values of P+R and P+A for the White, Chicano and Black examinee groups were 24 and 24, 17 and 18, and 14 and 15, respectively, with an omitting rate of four percent for each group; on item 52, the values were 36 and 72, 27 and 56, and 15 and 29, with omitting rates of 50, 52, and 49 percents, respectively. As a result of this tendency for similar

Table 12

Selected Guessing-Related Indices for Difficult GRE-Q Items, for White, Chicano, and Black Females

	White (N = 2000)				Chicano (N = 130)				Black (N = 1715)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
Data Interpretation (14 items)												
7	28	28	8.0	06	13	13	---	11	11	12	11.6	13
10	24	24	12.8	60	17	18	12.2	62	14	15	13.3	41
19	19	21	11.7	77	08	10	10.2	86	10	12	12.9	88
20	13	14	9.8	67	12	13	10.6	65	15	18	11.8	85
21	42	63	12.0	45	28	41	12.7	59	25	37	13.4	72
32	18	30	11.8	79	07	12	12.0	74	06	11	13.5	89
33	13	23	12.6	92	09	14	12.8	91	07	12	13.6	90
34	44	72	12.4	83	34	60	12.6	88	27	44	13.7	92
Algebra (10 items)												
27	54	57	11.4	77	38	40	12.7	79	25	27	12.5	85
39	38	41	12.0	66	19	21	11.4	81	13	15	12.5	80
41	61	79	12.0	84	48	59	13.6	87	30	35	15.0	69
53	16	26	13.7	97	10	12	13.8	90	11	14	14.9	98
55	05	05	----	76	04	04	----	82	04	04	----	92
Geometry (14 items)												
42	72	85	10.9	96	59	71	13.0	92	39	50	13.0	95
46	40	54	12.6	88	37	47	13.6	76	31	39	13.6	96
48	31	37	12.9	46	21	26	12.8	49	21	25	13.0	69
52	36	72	12.3	97	27	56	12.6	70	15	29	13.2	94
54	24	28	13.2	94	17	19	16.0	93	07	08	14.0	91
Arithmetic (12 items)												
45	28	45	12.6	93	19	31	13.4	94	16	23	14.2	98
49	31	36	12.2	53	24	27	13.9	72	29	33	13.5	87

Table 13

Selected Guessing-Related Indices for Difficult GRE-Q Items, for White Females Matched to Chicano and Black Females

	Matched to Chicanos (N = 390)				Matched to Blacks (N = 360)			
Item	P+R	P+A	MnO	RU	P+R	P+A	MnO	RU
Data Interpretation (14 items)								
7	18	18	11.3	06	10	10	13.3	13
10	14	14	15.6	52	11	11	13.4	66
19	11	14	11.7	75	06	08	12.9	86
20	08	09	10.6	78	06	07	11.9	80
21	29	50	12.8	72	21	34	13.4	73
32	09	16	12.5	82	08	16	13.0	87
33	10	18	13.0	86	11	18	13.4	86
34	35	61	12.8	86	36	58	13.8	92
Algebra (10 items)								
27	35	37	12.9	78	26	27	13.7	78
39	24	28	11.9	66	14	17	12.5	88
41	46	56	13.3	78	39	47	14.4	73
53	12	17	14.9	97	11	13	15.4	95
55	03	03	----	87	01	01	----	96
Geometry (14 items)								
42	57	72	11.3	91	43	55	13.0	98
46	29	38	13.4	85	30	37	14.3	96
48	25	28	12.9	48	20	23	14.2	54
52	30	59	12.5	93	16	41	12.8	92
54	16	19	13.5	89	12	13	13.1	94
Arithmetic (12 items)								
45	21	35	13.1	98	17	25	13.6	90
49	25	28	13.1	72	21	23	14.2	72

within-item omitting rates across examinee groups, the pattern of P+R values was closely approximated by that of the P+A values. Thus, the correspondence between the GRE-Q score pattern and the P+A pattern was similar to that between the score pattern and P+R values. As was true for male examinee data, shifts from P+R to P+A among GRE-Q items did not appear to be related to the item format distinction between Data Interpretation and Other Quantitative items.

Quantitative items, female examinees: MnO. There are 15 instances of Reverse Omitting indicated in Table 12 by underlined MnO values for Mno ≥ 13.5. Among these, eight were associated with Low P+R's ranging from eight to 17, another five with P+R's ranging from 19 to 31, and two with P+R's of 37 and 48. The correspondence of occurrences of Reverse Omitting to those of Low P+R values resembled similar data for GRE-Q male responses, and like the latter, the correspondence was not nearly as close as was true for both male and female responses to the GRE-V.

Using as the criterion of reversed omitting the value MnO > 13.0, the observed proportions of Low P+R items that were also reverse omits were 1/4 for Whites, 1/7 for Chicanos, and 7/12 for Blacks. A clearly higher proportion of Low P+R's associated with Reverse Omitting is characteristic of Blacks, for both males and females, for both GRE-V and GRE-Q items. Such a difference may be attributable to some aspect of ethnicity, or it may be an artifact of the fact that the Black examinees had lower overall scores, and were therefore encountering a larger proportion of difficult items. The MnO data in Table 13, for White females matched to Chicano and Black females, provide a test of these alternatives. If the differences in Reverse Omitting are primarily a function of ethnicity, then there should be a very low rate of Reverse Omitting for the two matched samples, in keeping with the low rate for the original White sample, since the matched groups of examinees were drawn from the same set of White GRE examinees. On the other hand, if the effect is primarily that of low-scoring examinees encountering a large enough proportion of difficult items to yield a Reverse Omitting effect, then the incidence of Reverse Omitting should correspond to the GRE-Q means of the five female groups involved.

In comparing the MnO data in Table 12 and 13, the hypothesis relating Reverse Omitting to group mean scores is supported. The GRE-Q mean scores were, in descending order of magnitude: Whites, 26; Chicanos, 18, and Chicano-matched Whites 18; Blacks, 14, and Black-matched Whites, 13. These means would predict Reverse Omitting frequencies going from low to high. Those observed were: 2; 6, 7; 12, and 13, the order predicted.

Summary of Phase I

Phase I derived three statistical indicators which could be useful in identifying differences in guessing behavior for ethnic and sex subgroups. The first indicator was an inordinately low level of success on the item, in the sense of proportion passing. This indicator was actually defined in two ways: P+R and P+A, depending on the base N for the proportions (these reaching vs. those responding). An arbitrary level, $P+ \leq 16$ was used to identify items with inordinately low levels. These percentages were "inordinate" or "dysfunctional" in the sense that the group could do better, on the average, through unconsidered, random responding. Thus, the indicator identifies items for which "guessing" is not successful for the group. The second indicator was a mean total test score for Omits higher than the average score for the group. This mean for the Omit group was labelled MnO, and an arbitrary level $MnO > 13.5$ was used to identify items which showed the phenomenon. The logic of the indicator is that more able people may be anticipated to do better, through guessing, than less able people, because they can eliminate more options. When the value of MnO rises, however, more able people are not guessing as frequently as less able people. For some analyses the criterion was relaxed to $MnO > 13.0$.

The third indicator was RU, the Pike and Flaugher modification of the Shannon information index. This indicator reflects evenness or rectangularity of distribution. It has been used in prior work as a general measure of randomness and hence of blind guessing.

A detailed inspection of these indices computed separately for six groups in an ethnic X sex analysis, for each item in the GRE and for the subsets of items, revealed only one finding of a potential difference between groups. This was that for Chicano females who omit, the average total score is lower than that of the matched group. In all other cases, no group differences were found.

The results of Phase I, then, clearly support the view that the standard instructions for the GRE are received in similar ways by the various ethnic groups, that the scoring formulas are equally appropriate for these groups, and that there are no differences in guessing behavior independent of differences in average level of ability.

The indicators applied in Phase I, however, are at best very general tests of implicit guessing behaviors. Each derives its relevance for group comparison from a logical relationship of "sensitivity to guessing." P+R and P+A may indicate poor guesses; MnO may indicate poor omissions; and RU may indicate differences in random-like behavior. These general tests lead to inferences of similar process but there is a need to test the inference itself. Phase II was an attempt to assess consistency of item reactions across the groups using analytical subcomponents of solution processes.

PHASE II STUDY

In Phase II, four informal and explanatory studies of item process were undertaken. The general spirit of the inquiry was an attempt to evaluate certain item-response strategies in terms of their potential for finding process differences among the groups. While the results of Phase I indicated general consistency of solution process with respect to the frequency of guessing and the strategies used, Phase II sought a deeper level of analysis. Accordingly, four special measures, based on item components, were developed. Candidate reports with respect to these four special measures were studied for consistency with the results of Phase I and for other inferences. These measure are described in the following sections.

Measures

The four kinds of supplementary data were each associated with a given item format. Thus, the Word Associations measure was administered to check on working hypotheses of how Analogies are answered, the Contextual Clues measure to see how well examinees might use immediate context to help answer Sentence Completion items; the measure of self-assessed Recognition Vocabulary as a source of information about answering Antonyms, (and, secondarily, Analogies and Sentence Completion items); and finally, a Quantitative Measure, focused primarily on mathematical terminology and symbols and on basic computational skills that are assumed to be known by GRE-Aptitude examinees.

Word associations measure. This measure was designed to obtain word association data related to analogy items, in order to investigate the possibility that word associations could have a major influence on analogy answering, particularly on the part of low-scoring examinees. Using the Relative Uncertainty procedure, Pike and Flaugher found that inner-city Blacks had a large random-like component in their responses to other verbal item formats in the PSAT, but responded much more systematically to analogy items, at the same time scoring comparatively worse. An inspection of Black male responses in the 1974 GRE d^ta suggests a similar situation. On 9 of the 17 Analogy items, the grou, selected the correct answer at less than the chance level of 20 percent; the median for the 9 items was only 8 percent pass. Something systematic would appear to be putting these candidates at a serious disadvantage in the way they tend to answer analogies.

The directions for solving GRE analogies may be deceptively simple for many examinees. They read as follows:

> In each of the following questions, a related pair of words or phrases is followed by five lettered pairs of words or phrases. Select the lettered pair which best expresses a relationship similar to that expressed in the original pair.

It would appear that examinees often respond as though the last sentence read: "Select the lettered pair that is most related to the original pair." This may lead, especially in instances where the relationship between stem words is not immediately evident, to answering on the basis of constrained word associations.

The problem of possible word-association effects in analogy scores is compounded by the possibility that these differ systematically by ethnic group membership. Campbell and Belcher (1975) compared free word associations of students at predominantly White and predominantly Black colleges, using stimulus words such as those appearing in the GRE. Some systematic differences were found, most of which seemed attributable to misreadings. For Chicanos the differences could be even more pronounced, because of bilingual effects.

Virtually any pair of words will fit into individual networks of word-association patterns in varied and complex ways. For the question at hand, free associations are not very productive, since only occasionally might a given word from a particular analogy item yield two or three free word associations which are among the other words used in the item. Even then, the relative strength of the free associations would not necessarily be the same as those operating within the contextual constraints of the analogy in question. For these reasons, constrained word associations were evoked, as follows. Consider the following analogy:

Song : Repertoire ::

(A) score : melody
(B) instrument : artist
(C) solo : chorus
(D) benediction : church
(E) suit : wardrobe

Assuming that the most important constrained word associations inappropriately influencing guessing behavior on such an item will be those among the left-hand terms and those among the right-hand terms, the item can be split into its left- and right-hand components. This serves the added purpose of keeping the word association responses from being contaminated by subjects inadvertently using the relationships between words of a given pair. Thus, for the above item, there were two items in the Word Association Measure shown in Appendix A. These are:

A. SONG

		Least Related			Most Related	
1.	score	1	2	3	4	5
2.	instrument	1	2	3	4	5
3.	solo	1	2	3	4	5
4.	benediction	1	2	3	4	5
5.	suit	1	2	3	4	5

J. REPERTOIRE

		Least Related			Most Related	
46.	melody	1	2	3	4	5
47.	artist	1	2	3	4	5
48.	chorus	1	2	3	4	5
49.	church	1	2	3	4	5
50.	wardrobe	1	2	3	4	5

Note that in the actual measure, the two sets of terms are separated by several intervening sets, to further insure that only the "vertical" associations are used.

Contextual clues. A potential component of partial information (PI) for all sentence completion items is that of short range contextual constraint. This can serve the examinee well if the short range context provides sufficient PI and if he or she uses the PI to eliminate implausible choices, but reserves judgment about making a final choice based on the alternative that best fits the immediate context. The latter precaution holds, because the answer that seems most plausible when only short range context is considered does not necessarily provide the best fit to the entire sentence. In order to test the stability and adequacy of examinees' use of short range context, the supplementary measure included questions calling for the rank ordering of choices when only limited context is provided. These questions are shown in Appendix A.

Recognition vocabulary. Vocabulary plays a particularly central role, of course, in answering antonyms. Omitting versus responding in the case of very difficult stem words is of interest, because in such instances most low-scoring examinees will have no rational basis for eliminating any of the alternatives. Examinees' answering behavior when very difficult words are used as distractors is also of interest. Because antonyms tend to be items of the "one clearly correct answer" type, a candidate who recognizes one of the alternatives as the opposite to the stem word can select that choice with confidence, even if one or more of the other alternatives is unknown to him.

Vocabulary limitations also pose a problem in guessing strategy on some of the more difficult analogy items. If a stem word is only vaguely known to some examinees, such as the word "repertoire" in the analogy example earlier, then those examinees would probably be well advised to omit the item, having no real basis for answering. Encountering analogy items with one or both words of a distractor that are unknown poses another problem, related to the tendency for analogies to be "best-answer," rather than "one right answer," in nature.

In order to examine the responses of the different subgroups of examinees to vocabularly material such as those described above with antonyms and analogies, selected words from such items in the 1974 GRE were included in the Recognition Vocabulary as part of the supplementary measure.

Quantitative measure. In the GRE-Q, a knowledge of certain mathematical terminology, symbols, and basic computational skills is assumed. These include "average," "successive integers," knowing the number of degrees in a right angle and in a circle, recognizing the symbols for representing inequalities, knowing how to solve simple inequalities, and so on. Where such information is assumed, but is not in fact known to some examinees incorrect answers presumably attributable to faulty quantitative reasoning may well be due in part to these specific gaps in quantitative information. From the standpoint of guessing strategies, it is evident that examinees lacking such information have little rational basis for making informed guesses when confronted with items requiring the basic information in order for the intended quantitative reasoning to be carried out. These considerations motivated the development and inclusion of the Quantitative Measure, shown in Appendix A.

Background information. In addition to the special measures, four items of background information were requested: sex, educational status, major field of study, and years since last attending high school. Ethnic group membership was already available because each of the four subject groups was homogeneous in that respect.

Methods

Instruments. The supplementary data were obtained through the use of two instruments. The first, the Control Test of Academic Ability (Peterson, 1965), is a 12-minute multiple-choice test of verbal and quantitative ability, designed for providing group indices of ability when more extensive data, such as SAT or GRE scores, are not readily available. The second, designed for the present study, consisted of five parts: "Background Information," "Word Association Measure," "Using Contextual Clues," "Recognition Vocabulary," and "Quantitative Measure." A copy of this instrument is attached, as Appendix A. Its contents are described in the previous section.

Subjects. The supplementary data were obtained from four *ad hoc* groups of paid volunteer college students: 94 Blacks at a major Southeastern Black college; 92 Chicanos, 12 at a state university and 80 at a community college in the Southwest; and 38 Blacks and 98 Whites at a state university in the Midwest. The percentages of males in the four groups were approximately 45, 55, 30, and 40, respectively. Means for Whites, Chicanos, and Blacks on the 30-item Control Test of Academic Ability (CTAA) were approximately 19, 13, and 14, with standard deviations of about 4. These values compare to State University, State College, and Junior College means of about 20, 18, and 15, and standard deviations again of about 4, as reported by Peterson (1965). It should be noted that the small sample sizes, differences in grade level and in percentage of males and females, and so on, make it quite inappropriate to consider the results as representive of the several ethnic and geographic groupings in any overall way. Rather, the data should be read for the purpose intended--that of describing the groups individually in order to better interpret the supplementary data regarding word associations, the ability to make use of contextual clues, and so on.

Data collection. At each of the four locations, data were collected in two sessions, with experimental subjects about equally divided between the two. This was done to keep the group sizes at a level that would allow careful monitoring of the data gathering sessions. About 75 minutes were needed to administer the two instruments. Each administration was personally conducted by the author with the help of collaborators at the respective institutions.

Word Associations

In studying the possible influence of word associations on the answering of verbal analogies, it must first be shown whether there are sufficient differences in the associative strengths among the choices in a given analogy for examinees to be subject to a systematic associative influence on their answering behavior. Secondly, the differences among analogy choices in their relative attractiveness on a purely analogical basis must be determined, to provide a differentiation between these two bases (associative and analogical) for responding. That is, the ratings of associative strength can be used to establish differences in relative associative attractiveness among the choices within items. If the presence of such intra-item characteristics is established the responses of selected examinee groups to the analogy items may be examined to determine whether these responses reflect susceptibility to the potential associative influence inherent in each item. However, the role of analogical attractiveness confounds the assessment of associative strength as a determiner of response.

In the present study, this confounding could not be fully analyzed. To explore the relationship between item properties and item difficulty, measures of each of the two potential answer tendencies were used.

Constrained word associations. Relative constrained word association values for selected analogy items are provided in Table 14. They are derived from responses of the four groups of students who completed the Word Association Measure shown in Appendix A. As noted earlier, each of the word-association "items" was made up of either the left- or the right-hand set of words from one of the analogy items. In the table, the word clusters in the "items" have been regrouped, so that words from the same stem pair (e.g., CISTERN and WATER) are adjacent, and are in the order to be used in subsequent tables.

For each subject group, the average of the five word association means involving a given cluster of stimulus words is necessarily 3.0. The potential range of mean values within a set of five is from a low of 1.0 to a high of 5.0. The observed range varies with both the cluster of stimulus words and the student group. For example, means of White students' associative rankings involving CISTERN ranged from 2.07 for "official" to 3.76 for "vault;" Chicano rankings of the same five word pairs ranged from 2.43 for "shower" to 3.63 for "museum." Word association means involving the stem WATER were more sharply differentiated within student group, and more highly consistent between groups. The weakest association to WATER was "antiques," with a mean of 1.26 for Whites and of 1.36 for Chicanos; the strongest association was to "cloud," with means of 4.66 and 4.58, respectively.

An inspection of all of the sets of constrained work association means suggests a clear-cut differentiation in associative strength within each set, particularly when the stem word is a familiar one,

Table 14

Means of Constrained Word Association Values, by Student Group

Stimulus words	Ad hoc student group				Overal mean
	White (N=98)	Chicano (N=92)	SE Black (N=94)	MW Black (N=38)	
K. CISTERN					
51 shower	3.19	2.43	2.60	2.55	2.73
52. official	2.07	2.67	2.49	2.91	2.46
53. science	2.78	3.16	3.01	3.39	3.03
54. museum	3.18	3.63	3.39	3.21	3.37
55. vault	3.76	3.10	3.52	2.91	3.40
C. WATER					
11. cloud	4.66	4.58	4.62	4.62	4.62
12. power	3.82	3.70	3.54	3.57	3.67
13. matter	3.33	3.07	3.30	3.41	3.26
14. antiques	1.26	1.36	1.46	1.46	1.37
15. valuables	1.94	2.27	2.09	1.95	2.08
A. SONG					
1. score	3.28	3.03	2.69	3.11	3.02
2. instrument	3.73	4.07	3.79	3.83	3.86
3. solo	4.40	4.04	4.70	4.56	4.40
4. benediction	2.29	2.34	2.50	2.19	2.35
5. suit	1.31	1.43	1.32	1.39	1.36
J. REPERTOIRE					
46. melody	3.46	3.40	3.43	2.97	3.38
47. artist	3.29	3.35	3.00	2.97	3.18
48. chorus	3.38	3.55	3.64	3.77	3.55
49. church	2.11	2.44	2.43	2.66	2.36
50. wardrobe	2.75	2.26	2.49	2.63	2.52
N. ANACHRONISM					
66. atheism	2.39	2.86	2.81	2.77	2.69
67. fallacy	3.53	3.32	3.43	3.16	3.40
68. propaganda	3.72	3.41	3.57	3.30	3.54
69. artifact	2.38	2.62	2.67	2.93	2.60
70. criterion	2.97	2.79	2.45	2.81	2.75
E. HISTORIAN					
21. skeptic	1.83	1.92	1.89	2.08	1.90
22. logician	3.09	2.95	2.79	3.22	2.98
23. politician	2.88	2.69	2.76	2.86	2.79
24. archaeologist	4.47	4.34	4.64	4.17	4.45
25. statistician	2.73	3.09	2.92	2.70	2.88

(continued)

Table 14--continued

Stimulus words	Ad hoc student group White (N=98)	 Chicano (N=92)	 SE Black (N=94)	 NW Black (N=38)	Overall mean
P. INVIDIOUS					
76. xxxxx	2.77	2.72	2.52	2.33	2.63
77. xxxxx	2.90	2.97	2.82	2.87	2.89
78. xxxxx	2.96	3.03	3.14	2.93	3.03
79. xxxxx	2.94	2.91	2.92	2.67	2.89
80. xxxxx	3.43	3.37	3.55	4.17	3.54
M. ODIUM					
61. xxxxx	1.77	2.49	2.49	2.25	2.24
62. xxxxx	3.67	3.29	3.17	2.97	3.33
63. xxxxx	3.51	3.18	3.23	3.50	3.33
64. xxxxx	3.09	3.06	3.06	3.25	3.09
65. xxxxx	2.95	2.98	2.98	3.03	2.98
O. DIFFIDENT					
71. xxxxx	2.31	2.85	3.05	3.00	2.76
72. xxxxx	2.74	2.74	2.79	3.00	2.79
73. xxxxx	3.31	3.32	3.36	3.07	3.30
74. xxxxx	4.02	3.88	3.62	3.57	3.81
75. xxxxx	2.62	2.22	2.19	2.34	2.45
G. CONFIDENCE					
31. xxxxx	2.65	2.96	2.48	2.63	2.69
32. xxxxx	1.90	1.89	1.95	1.91	1.91
33. xxxxx	3.74	3.76	3.82	3.76	3.77
34. xxxxx	3.04	2.69	2.94	2.71	2.87
35. xxxxx	3.68	3.73	3.81	4.03	3.77
B. SKUNK					
6. camel	1.37	1.77	1.52	1.44	1.54
7. porcupine	4.40	4.21	4.22	4.31	4.28
8. lion	2.49	2.67	2.66	2.39	2.58
9. cat	3.96	3.47	3.88	3.61	3.76
10. hound	2.77	2.88	2.71	3.25	2.84
H. SCENT					
36. hump	1.87	1.86	2.09	2.03	1.95
37. quill	2.84	2.71	2.78	2.68	2.77
38. mane	3.26	3.19	3.29	3.32	3.26
39. whisker	3.82	3.58	3.83	3.92	3.77
40. ear	3.22	3.65	2.99	3.05	3.26

(continued)

Table 14--continued

Stimulus words	Ad hoc student group				Overall mean
	White (N=98)	Chicano (N=92)	SE Black (N=94)	NW Black (N=38)	
F. OBSESSION					
26. emotion	4.40	4.43	4.31	4.50	4.39
27. author	1.55	1.58	1.71	1.74	1.63
28. experimentation	2.17	2.41	2.72	2.42	2.43
29. thought	3.57	3.75	3.68	3.92	3.69
30. vigil	3.32	2.84	2.59	2.42	2.86
D. IDEA					
16. thought	4.79	4.66	4.83	4.66	4.75
17. play	1.33	1.52	1.37	1.51	1.42
18. theory	3.63	3.97	3.68	3.60	3.74
19. dream	3.31	2.84	3.30	3.23	3.16
20. attention	1.95	2.01	1.83	2.11	1.95
I. SUPPLICATE					
41. xxxxx	2.58	2.45	2.53	2.77	2.55
42. xxxxx	3.56	3.79	3.73	3.31	3.65
43. xxxxx	3.29	3.49	3.59	3.51	3.46
44. xxxxx	2.66	2.70	2.29	2.57	2.55
45. xxxxx	2.91	2.54	2.86	2.83	2.78
L. HUMBLE					
56. xxxxx	4.75	4.55	4.70	4.59	4.66
57. xxxxx	2.52	2.76	2.56	2.59	2.61
58. xxxxx	3.23	2.67	2.90	3.27	2.98
59. xxxxx	2.39	2.64	2.53	2.73	2.54
60. xxxxx	2.12	2.38	2.33	1.82	2.22

Note: In order to maintain the security of test items still in use, the words accompanying stimulus words P, M, O, G, I, and L have been deleted.

such as WATER, SONG, or HISTORIAN. Clearly, the groups perceive distinctions to the task, and report differentiation in association.

There is a general consistency in the ranking of word association means for each cluster of words across the four student groups. This consistency is apparently strongest in instances where the stem word appears to be well-known. For the cluster of associations involving a rare word like CISTERN, on the other hand, differences in relative associative strength interact with student group. The word "vault," for example, has the highest association values with CISTERN for White and SE Black students, but middle-level ones for Chicanos and MW Blacks. Similarly, "official" tends to be seen as least associated by Whites and SE Blacks, but to have a middle level of relationship to CISTERN for the other two student groups. With that exception, weaker correspondence in word association rankings across student groups for difficult stem words appears to be a secondary effect, resulting from a compression in the range of word association means across the five sets in a cluster. Representative of this effect are the means involving the stem word ANACHRONISM. For each of the four student groups, the five words related to ANACHRONISM fall into essentially two levels: associative means for "atheism," "artifact," and "criterion" are consistently lower (2.38 to 2.97), with little differentiation among them, and those for "fallacy" and "propaganda" consistently higher (3.16 to 3.72).

In general, then, the findings of the word association measure indicate a marked comparability among the groups. While some differences exist, the approach of the groups to this specialized task calling for dealing with item components, is in the main highly similar. The implication would be that the application of association in response would be similar for the groups.

Because all 16 stem words used in the above analyses were included in the Recognition Vocabulary measure, information from the latter was used to see whether the differences in the differention among mean association values within student group for each cluster, and the consistency across student groups, are related to the recognition level of the stem words. Five of the 16 words had low recognition values for all four student groups. In the order of their appearance on Table 14, these are: CISTERN, ANACHRONISM, INVIDIOUS, ODIUM, and DIFFIDENT. A sixth word, REPERTOIRE, was marginally unfamiliar to Chicanos and Southeast Blacks. All other stem words were well-known to all four groups of students. In general, recognition level is related to association values.

Table 15 reports the correlations of association across groups. These data provide evidence of considerable comparability in constrained word associations, sufficient to warrant averaging

Table 15

Correlations between Student Groups' Mean Constrained Word Associations, for Items Grouped by Recognizability of the Stem Word

Ad hoc subject group	N	Ad hoc subject group White	Chicano	SE Black	MW Black
White	98	--	96	97	96
Chicano	92	73	--	96	95
SE Black	94	76	91	--	97
MW Black	38	55	76	76	--

Note: Correlations above the diagonal are for the 55 word-associations involving the 11 high-recognition stem words; those below the diagonal are for the 25 associations involving the 5 low-recognition stem words.

across groups to get overall mean values for the word associations. At the same time, these data provide strong confirmation of the relationship between recognizability of the stem word and between-group consistency in constrained work associations. Between-group correlations for word associations involving the 11 familiar stem words were all .95 or higher. Of those for associations involving the five relatively unfamiliar stem words, the lowest was .55, between Whites and Midwest Blacks, and the highest was .91, between Chicanos and Southeast Blacks. The other correlations involving the unfamiliar stem words were in the .70's.

The overall means provided in Table 14 can be examined for evidence bearing on the relationship between stem-word familiarity and the degree of differentiation between weakest and strongest constrained word associations. The maximum possible range between association means within a cluster is 4.0 (one word rated 5.0, another rated 1.0). For the stem-word CISTERN, the range in overall means was 3.40 - 2.46, or 0.94. Ranges for the other difficult stem words, ANACHRONISM, INVIDIOUS, ODIUM, and DIFFIDENT, were 0.94, 0.91, and 1.36, respectively. Ranges for two of the other stem words were of similar magnitude (1.19 for REPERTOIRE and 1.10 for SUPPLICATE), but those for the remaining stem words were substantially higher, themselves ranging from 1.82 for SCENT, to 3.25 for WATER, and 3.33 for IDEA. It is clear, then, that there was a strong tendency for lower-recognition stem words to yield less differentiation in word association strength than did the higher-recognition stem words.

Distracter characteristics. Indices of associational strength and analogical strength characteristics are presented in Table 16. These are 1) average constrained word associations ratings and 2) judged analogical strength. Each value of the Associational Rating in Table 16 is the simple average of the values for the respective item components. Thus the associative rating for Choice A of Item 1 in Table 16 is the average of the associative rating for CISTERN - shower (2.73) and that for WATER - cloud (4.62); i.e., 3.67. (The separate associative ratings are given in the "Overall mean" column of the Table 14 data.) This method of combining associative ratings seems appropriately sensitive to the degree of differentiation in associative strength among the five choices presented with each of the stem words. Thus, the associative ratings involving WATER, which had a range of 3.25, will contribute relatively more to differences in the combined associative ratings than will the ratings involving CISTERN, which had a range of only 0.94.

The range of these average associative ratings across the choices for each analogy item is given in the parenthetic entry

Table 16

Choice Attraction Indices Derived from Supplementary Data, and Response Percentages for Data-Tape Groups for Selected Analogy Items

			Response percentages for Data-Tape groups					
	Suppl.	Judged	Males			Females		
Response choice	Assoc. rating	analog. rating	Wh	Ch	Bl	Wh	Ch	Bl
1. CISTERN : WATER								
Omit								
A shower : cloud	3.67	1.25	(No response data for this item.)					
B official : power	3.07	3.00						
C science : matter	3.14	1.75						
D museum : antiques	2.37	4.00						
*E vault : valuables	2.74	5.00						
(Choice differentiation)	(1.3)	(3.8)						
2. SONG : REPERTOIRE								
Omit								
A score : melody	3.20	1.25	(No response data for this item.)					
B instrument : artist	3.52	2.25						
C solo : chorus	4.00	3.00						
D benediction : church	2.36	3.50						
*E suit : wardrobe	1.94	5.00						
(Choice differentiation)	(2.1)	(3.8)						

(continued)

Table 16 - continued

Response choice	Choice attraction index		Response percentages for Data-Tape Groups					
	Suppl.Ss' assoc. rating	Judged analog. rating	Males			Females		
			Wh	Ch	Bl	Wh	Ch	Bl
3. ANACHRONISM : HISTORIAN								
Omit			16	20	17	13	18	17
A atheism : skeptic	2.30	1.25	2	3	4	2	4	6
*B fallacy : logician	3.19	5.00	47	30	21	48	24	18
C propaganda : politician	3.16	3.25	6	11	22	9	15	25
D artifact : archaeologist	3.51	3.50	25	26	27	22	29	27
E criterion : statistician	2.82	2.00	5	11	8	5	11	6
(choice differentiation)	(1.2)	(3.8)						
(Item associative response tendency)			(3.2)	(3.6)	(3.7)	(3.2)	(3.7)	(3.8)
4. INVIDIOUS : ODIUM								
Omit			6	52	49	58	50	46
*A xxxxx : xxxxx	2.44	5.00	14	7	7	20	11	7
B xxxxx : xxxxx	3.11	1.25	10	12	14	10	15	9
C xxxxx : xxxxx	3.18	3.25	6	12	10	5	10	18
D xxxxx : xxxxx	2.99	3.50	5	8	7	3	6	10
E xxxxx : xxxxx	3.26	2.00	4	9	13	4	9	9
(choice differentiation)	(0.8)	(3.8)						
(Item associative response tendency)			(3.8)	(4.1)	(4.2)	(3.7)	(4.1)	(4.0)
5. DIFFIDENT : CONFIDENCE								
Omit			36	35	20	29	25	23
A xxxxx : xxxxx	2.72	3.00	10	11	17	12	18	15
*B xxxxx : xxxxx	2.35	5.00	10	6	18	11	10	4
C xxxxx : xxxxx	3.54	3.50	9	12	16	11	14	22
D xxxxx : xxxxx	3.34	2.50	18	20	21	20	15	17
E xxxxx : xxxxx	3.06	1.00	17	16	17	16	18	19
(choice differentiation)	(1.2)	(4.0)						
(Item associative response tendency)			(4.2)	(4.3)	(4.3)	(4.2)	(4.2)	(4.4)

(continued)

Table 16 - continued

	Choice attraction index		Response percentages for Data-Tape Groups					
	Suppl.Ss' assoc. rating	Judged analog. rating	Males			Females		
Response choice			WH	Ch	Bl	Wh	Ch	Bl
6. SKUNK : SCENT								
Omit			3	4	5	2	5	9
A camel : hump	1.74	3.50	5	9	17	5	10	21
*B porcupine : quill	3.52	5.00	82	68	59	84	65	51
C lion : mane	2.92	2.25	1	7	4	0	4	3
D cat : whisker	3.76	2.00	1	3	3	1	2	5
E hound : ear	3.05	2.25	9	10	12	8	15	12
(choice differentiation)	(2.0)	(3.0)						
(Item associative response tendency			(2.8)	(3.0)	(3.0)	(2.8)	(3.0)	(3.1)
7. OBSESSION : IDEA								
Omit			8	13	14	7	11	13
A emotion: thought	4.57	3.25	31	40	27	24	35	31
B author : play	1.52	1.25	1	3	4	0	1	3
C experimentation : theory	3.08	2.50	12	15	23	15	18	21
D thought : dream	3.43	3.25	5	12	13	5	10	18
*E vigil : attention	2.41	4.75	43	16	19	49	26	14
(choice differentiation)	(3.1)	(3.5)						
(Item association response tendency)			(3.5)	(4.3)	(4.1)	(3.3)	(4.0)	(4.2)
8. SUPPLICATE : HUMBLE								
Omit			38	36	39	31	36	40
A xxxxx : xxxxx	3.60	2.50	20	20	28	24	22	22
B xxxxx : xxxxx	3.13	3.25	24	20	14	23	24	12
*C xxxxx : xxxxx	3.22	4.75	7	10	5	11	7	9
D xxxxx : xxxxx	2.55	1.25	5	5	4	4	3	8
E xxxxx : xxxxx	2.50	3.25	6	10	10	7	8	8
(choice differentiation)	(1.1)	(3.5)						
(Item association response tendency)			(4.1)	(4.0)	(4.2)	(4.1)	(4.1)	(4.1)

called choice differentiation. The theoretical maximum differentiation is again 4.0. But the possible maximum is necessarily limited to the average of the differentiations for the two component associative ratings, and the possible maximum is achieved only when the independent "halves" of the association task agree. Thus, for item 7, OBSESSION : IDEA, the Choice A words were each highest in associative attraction, and the Choice B words were each lowest. The resulting choice differentiation of 3.1 was the largest one observed among the eight items. Item 1, CISTERN : WATER, provides a counter example. There the choice differentiation of 3.25 for associations to WATER, averaged with a choice differentiation of 0.94 for associations to CISTERN, implies a theoretical maximum choice difference of 2.1 for the combined associative rating. However, the high combined associative rating observed for Choice A (3.67) is derived from a high value for WATER-cloud of 4.62 but a low value for CISTERN-official of only 2.73. Similarly, the low combined associative rating observed for Choice D is derived from a very low association to WATER combined with a moderately high one to CISTERN, for an average of 2.37. The resulting choice differentiation is associative rating for Item 1 was just 1.3, substantially lower than the possible 2.1.

The choice differentiations for Associative Ratings are sufficiently large to demonstrate a potential for a large word association factor in examinees' responses to analogy items. Since the maximum theoretical choice differentiation is 4.0, the observed minimum, median, and maximum values for the eight analogies of 0.80, 1.25, and 3.10 may also be considered as 20, 21, and 77 percent, respectively, of that maximum.

For the present study, a priori ratings of analogical values among the choices for each item were developed as a supplement to the associative indicator. Useful a priori ratings of analogical values among item choices could by no means be assumed. Two criteria would have to be met: (1) there would have to be good inter-judge agreement on such ratings, and (2) there would have to be sufficient differences in analogical value or goodness across the five choices for each item to give a reasonably high discrimination level among the choices.

Four judges--the author and three other persons familiar with writing and editing analogy items--provided the ratings, using the instructions shown in Appendix B. Here, of course, items were presented in their entirety, rather than splitting each one into its left-hand and right-hand components, as had to be done for obtaining word association values.

The first criterion for the usefulness of the Judged Analogical Ratings, that of good inter-judge agreement, was well met. The average correlation between individual judges was r = .70; the Spearman-Brown reliability index for judgments averaged over the four judges was r = .90. As indicated by the Judged Analogical Rating column in Table 16, the second criterion, that of adequate discrimination in analogical ratings across the five choices for each item, was also well met. Just as for the Students' Associative Ratings, the theoretical maximum differentiation between highest and lowest ratings was 4.0. For the Judges' Analogical Ratings, the median Choice Differentiation value was 3.8, or 95 percent of the theoretical maximum.

The judged analogical ratings among the five choices for each item provide a measure of choice attraction based on inherent logical reasonableness. For examinees properly understanding and following the analogies directions, these ratings should reflect the attractiveness of the several choices on an analogical basis. Those fully able to solve a given analogy item would, of course, select the correct response, which would have the highest analogical value. Those next to the highest in the ability to solve the same analogy would presumably be selecting between the best two choices, unable to definitely identify the correct choice because of the subtleties in relationships, or vocabulary limitations, etc., but able to eliminate the weaker choices--i.e., the choices with low analogical values.

The data in Table 16 also include, for six items, actual item analyses data for Whites, Chicanos and Blacks separately by sex. These data were derived from the sample of December 1974 test candidates. A fundamental question in the application of indices such as the ratings of associational and analogical strength is their actual relationship to candidate behavior. In attempting an assessment of this relationship, the first decision was to eliminate the key and to consider the properties of the wrong answers only. The second decision combined the sexes by averaging the percentages reported in Table 16. Then the four distractors for each of the six items were systematically ranked on five criteria: associational ratings, analogical ratings, and relative "popularity" for Whites, Chicanos and Blacks, respectively. Values of Kendall's Tau were computed as measures of association between ranks. Table 17 presents these values.

This table shows that Items 3 and 7, the middle difficulty items, have the strongest relationship on both indices, while item 4 is the most poorly predicted item. In general, the two indices tend to show rough similarity of values, so that either might explain the popularity rankings of the item data. It is plausible to interpret the findings as supportive of modest associational strength for each index. The best test of overall significance for these data

Table 17

Relationship of Distractor Association Strength and Distractor Analogical Strength to Proportion Selecting this Distractor

Association Strength

	Ad hoc subject group		
Item	White	Chicano	Black
3	+1.00	+1.00	+1.00
4	-0.18	.00	+ .33
5	.00	.00	+ .91
6	.00	- .33	- .33
7	+ .66	+ .66	+ .66
8	+ .33	+ .33	+ .66

Analogical Strength

	Ad hoc subject group		
Item	White	Chicano	Black
3	+1.00	+1.00	+1.00
4	-0.55	-0.66	-0.33
5	-0.66	-0.66	-0.55
6	-0.18	+0.55	+0.55
7	+0.55	+0.55	+0.55
8	+0.55	+0.55	+0.18

All entries are values of Kendall's Tau. 11 of 18 entries are positive non-zero, a pattern significant at the .05 level (p=.036).

seemed to be the exact probability computation of the number of positive nonzero values they contain. For rankings of four things, a positive nonzero value of Tau is achievable by chance with a probability of .375. Thus, the 11 of 18 positive nonzero values for each indicator yield an overall finding of significance at about the .05 level.

The roughness of these processes is indicated in the linking of ratings by small *ad hoc* samples to item analyses findings in quite different groups, and this roughness implies a severe test for the assessment of relationship. This should be borne in mind in assessing the modest results reported; further work with item component strategies seems indicated by these results.

Contextual Clues

The Contextual Clues measure consisted essentially of truncated sentence completion items. The truncation consisted of the deletion of material before and/or after the stimulus blank in the stem. This deletion shortened the range of verbal stimuli which could be used to generate an answer. The task, then, offers information on a response component calling for an inferential process which is not logically different in kind from the full-item process which the item type demands. Put another way, this says that the fundamental cognitive task required by the contextual clue items is very similar to the task that the untruncated items demand. There is less basis for asserting that there is a correct answer, one logically compelled by the references in the stem and the general canons of language usage, but the task is basically the same: select the response which, in your judgment, best restores the intended meaning.

Table 18 presents the mean ratings for the various responses for the various groups. The clearest patterns concern the "key," the answer which is most attractive. For two items, B and E, the groups tend to indicate the presence of a basically unequivocal "key," and to agree among themselves on what it is. The average rating strength for these "keys" across the groups differs very little. In B, the means vary from 4.01 to 4.36; in E, from 4.29 to 4.45. Item E offers a most plausible distracter, in number 24, "obsequious;" the range of its values across groups is quite small: 2.92 to 3.22.

Item D almost reaches this level of intergroup agreement. Only the deviation by the MW Black group on "manifest" (#20) breaks the pattern. Whereas the other three groups found "manifest" a clear second choice, MW Blacks made it their "key."

Table 18

Means of Contextual-Clue Response Ratings, by Subject Group

	Ad hoc subject group				
extual-clue item	White (N=98)	Chicano (N=92)	SE Black (N=94)	MW Black (N=38)	Overall mean
..growth is not a _____ rocess for all people...					
. uniform	4.28	3.19	3.72	3.49	3.71
. healthy	2.12	2.31	2.46	2.62	2.33
. unique	2.55	2.65	2.40	2.43	2.52
. simple	3.36	3.68	3.38	3.59	3.48
. progressive	2.66	3.17	3.01	2.86	2.93
..problems that are difficult oday may become _____ tomorrow.					
. insoluble	2.97	2.94	2.39	2.67	2.76
. manageable	4.27	4.01	4.18	4.36	4.18
. dissipated	2.59	2.46	2.53	2.44	2.52
. prominent	2.56	2.81	2.84	2.83	2.75
. vital	2.61	2.78	3.01	2.69	2.78
..pervasive feature of human ntellect is its _____ capacity...					
. finite	2.49	2.60	2.45	2.76	2.54
. inadequate	1.92	2.17	1.90	1.97	1.99
. remarkable	3.93	3.78	3.96	3.73	3.87
. boundless	3.94	3.80	3.91	3.81	3.88
. limited	2.72	2.64	2.83	2.75	2.73
..geese _____ no territorial ehavior.					
. identify	3.16	3.09	3.01	3.62	3.15
. merit	2.25	2.25	2.17	2.47	2.25
. divulge	2.80	2.60	2.59	2.83	2.69
. accept	2.95	2.89	2.98	2.69	2.80
. manifest	3.84	4.16	4.23	3.43	4.00
..he was famous for his ____ and irresponsible ehavior...					
bizarre	4.45	4.36	4.29	4.40	4.37
penurious	2.45	2.79	2.69	2.40	2.61
cavalier	2.33	2.40	2.24	2.23	2.31
obsequious	2.92	3.08	3.02	3.23	3.03
licentious	2.85	2.34	2.74	2.74	2.66

Items A and C showed split-"key" responses. On Item C this was found for all groups, and in a similar manner. Neither #13 nor #14 were distinguishable in this context. Item A showed Whites preferring #1, "uniform" to #4, "simple," and clearly. Only the SE Blacks mirrored this split, and then in a more muted manner. On Item A, also, the minority groups seemed to rate "progressive" as more plausible than the White group did.

The homogeneity of these results across groups is interesting. There are known differences in levels of success on verbal material for the various groups. Such differences could be interpreted as based upon solution-process problems, problems which would be exacerbated by a short range task of the type here posed. That is, the basic inferential task which the sentence completion items and the contextual clues pose is one of deducing a proper fit, of combining knowledge of the world and knowledge of language to reach a proper replacement from among the alternatives. Knowing that ethnic groups differ in the item task, one would anticipate their differing in the short-range task. Yet, in general, there do not seem to be differences which might explain item-success differences. Only Item A offers such a pattern among these items.

Recognition Vocabulary

The established differences in the success of the ethnic and racial groups on verbal material is plausibly due to a variety of factors: imbalances in the quantity or quality of exposure to language, or interferences in thought and learning due to association systems derived from competing language systems such as Spanish or Black English. The exploration of such complex hypotheses was clearly beyond this study, though obviously relevant to the interpretation of results. As a basic effort to develop information broadly useful in the analytic study of the item-types, the subjects were asked to supply self-ratings of word recognition.

These word-recognition values have been discussed earlier in the context of the analogy items. As stated there, some of the words were derived from the test content of the analogies, and they are useful in interpreting the results of that "item decomposition" study. The word recognition values had potential relevance also for the antonym items, which consist simply of single-word stimuli, as the stems of items, and primarily single words as distracters. While the sentence completion items offer more basis for deriving meaning through inference from context, recognition differences were plausibly a potential source of explanation for this item type, as well.

There were 69 words in the recognition task. The selection principle was test-centered, in the sense that the specific form of the GRE which figured in the data-tape analysis of Phase I was reviewed for possible words to use. Words were selected from any of the three item types: antonyms, analogies, and sentence completions. Further, words derived from either stem or responses to items.

The rating scale ranged from very unfamiliar (1) to very familiar (7), with a midpoint of 4.0. In the analysis, attention was focused on words which were relatively unfamiliar, which was arbitrarily defined as any word with a mean of less than 5.00 for a particular group.

Seventeen of the 69 words were judged relatively unfamiliar by this criterion for all of the four groups. These and 11 others were judged relatively unfamiliar by at least one of the groups. Table 19 presents the means for these 28 words. These values obviously reflect both differences in the subjectively imposed interpretation of the rating scale and actual familiarity with the words. The Chicano group tends to give the lowest rating for the 28 words, the MW Black group the highest, although this group is not very different from the White.

Table 20 appraises the consistency of ratings for the four groups, for the 69 words in the total list and for the 28 words in the "relatively unfamiliar" list. For the total list, the correlations are very high. The distributions are markedly skew, however, so that the coefficients could be more reflecting agreement as to which specific words are unfamiliar than agreement on the level of unfamiliarity for these words. This latter aspect is better reflected in the below-diagonal correlations, based on the 28 "unfamiliar" words. The high average level of these coefficients indicates that on the whole the groups did reflect similar relative judgments. While SE Blacks and the Chicano group seem most different, their correlation of .82 is still substantial. "Estrange," "conciliate" and "odium" constitute words which are judged relatively easier by Chicanos vis-a-vis SE Blacks, while "antecedent," "ecclesiastical," "hyperboles" and "proliferation" are judged relatively easier by SE Blacks vis-a-vis Chicanos. No clear rationale for these specific findings emerges. Nor are there any obvious potential content-based interpretations of other intergroup differences for specific words. "Plenum" and "hiatuses" are the most unfamiliar words of all.

These high intergroup correlations have some bearing on the sources of known group differences in item successes. The more general kinds of information reflected in the recognition levels do not show knowledge of words, which test items require, but they do

Table 19

Mean Recognition level of selected vocabulary i ems, by subject group

Vocabulary Item		Ad hoc subject group			
		White (N=98)	Chicano (N=92)	SE Black (N=94)	MW Black (N=38)
3.	acridness	3.49	2.69	2.87	3.34
5.	anachronism	4.00	3.15	3.65	4.47
8.	antecedent	5.66	4.48	5.62	5.81
10.	cistern	4.83	2.51	3.01	3.45
12.	conciliate	4.41	4.73	4.68	5.50
15.	corroborate	5.16	4.52	5.14	5.84
17.	diffident	4.15	3.87	4.32	4.87
22.	ecclesiastical	4.61	3.16	4.45	4.24
23.	estrange	5.17	4.85	4.43	5.86
24.	exigencies	2.59	2.54	2.54	3.41
25.	expedience	5.30	4.42	5.17	5.58
28.	heinous	2.96	2.79	2.99	3.19
29.	hiatuses	2.12	1.90	2.37	2.53
31.	hominid	3.05	2.61	3.12	3.66
33.	hyperboles	5.74	3.16	4.67	4.78
36.	incommodities	4.78	4.89	5.23	5.58
40.	invidious	3.12	3.23	3.48	4.16
44.	odium	3.01	2.87	2.63	3.26
46.	paucity	2.77	2.25	2.92	3.26
47.	placate	4.04	2.99	3.54	3.86
48.	plenum	2.21	1.87	2.00	2.81
50.	progenitorship	3.19	2.37	2.70	3.19
51.	proliferation	5.20	3.52	4.68	4.57
52.	promulgate	3.27	2.76	3.36	3.28
54.	repertoire	6.16	4.[illegible]	5.11	6.05
58.	secular	6.23	4.88	5.79	6.05
64.	surfeit	3.19	2.65	3.12	3.19
66.	tenacity	5.07	4.69	4.83	5.08

Note: Recognition levels are self-ratings of word recognition, on a 7 point scale ranging from 1 (words with which you are completely unfamiliar) to 7 (words with which you are very familiar, and have no doubt regarding their meaning).

Words for which the mean rating was 5.00 or greater for all four groups are not included in this list.

Table 20

Correlations between Subject Groups' Mean Recognition Levels for Selected Vocabulary Items

Ad hoc subject group	N	Ad hoc subject group			
		White	Chicano	SE Black	MW Black
White	98	--	94	97	96
Chicano	92	97	--	97	97
SE Black	94	95	82	--	98
MW Black	38	95	91	91	--

Note: Correlations above the diagonal are for the full set of 69 items. Those below the diagonal are for the subset of 28 Vocabulary Recognition items receiving an average recognition rating below 5.00 from at least one of the four subject groups.

indicate experience with the words. Within the limits of the study, they lend support to the notion that in a very broad sense, the exposure to language for these groups has rough equivalences. Translated into practical consequences for test developers, this means that there is little evidence of a specialized subvocabulary of words which are intrinsically biased against some one group or the other. If there are domains of vocabulary which are inappropriate for testing for various groups, they are not revealed in this limited analysis.

Quantitative Measure

The fundamental hypothesis underlying the exploratory research in the present study is the capacity to isolate component activities associated with item response. In the case of the verbal materials, such component activities retained the structure of multiple-choice questions. That is, the tasks in the Word Associations, Contextual Clues, and Recognition Vocabulary materials all involved a selected response from a predefined set. All entailed recognition processing.

The quantitative measure departed somewhat from this model. In all, there were 29 quantitative tasks, of which 24 were "free response" questions. Thus, item 1 in Appendix A calls for a constructed response to the task "3(5-6) = ?" The subjects had to perform the calculations called for, with no guiding options present.

Such data, then, are of interest in several ways. First, there is the basic attention to the level of correct response, the percentage of a group that can generate the correct answer. But it is also useful to consider what the actual erroneous response may be if someone cannot perform the indicated operations. The summary tables, therefore, reflect both of these kinds of information. Table 21 compares the four ethnic groups in terms of overall level of success. Appendix A shows the distributions of the various kinds of errors which were made.

A brief inspection of the kinds of mathematical operations which are demanded by the quantitative measure should be sufficient to establish that they involve only very basic concepts expressed in very simple numbers. Most numbers, in fact, are single-digit. The "difficulty" in such tasks, then, lies not in any intrinsic logical or mathematical subtlety, but more in what might be called a requirement of basic mathematical literacy. If one is to process swiftly and effectively the kinds of questions which are included in GRE-Q, one needs to command a ready interpretation of the basic operations which are sampled here.

ERIC Full Text Provided by ERIC

The content sampling is a curriculum-centered one, but a test-centered one. The particular 29 tasks which comprise the quantitative measure were chosen because they reflected the basic operations which were found in the specific GRE form which was chosen as the focus for this study. Some other GRE form might yield a somewhat different set of basic operations; a group of mathematics educators would almost certainly develop a different balance of samplings. In the context of the present study, however, the informal and test-centered approach seemed appropriate. If the method of "component analysis" used here seemed to give meaningful results, a more systematic exploration of a wider array of material was clearly possible.

Table 21 reveals that the <u>ad hoc</u> Chicano group, on the average, experienced the greatest difficulty, and demonstrated the least command of these basic operations. The only single task for which the <u>ad hoc</u> Chicano group was not the least successful was for the multiple-choice task involving the identification of an isosceles triangle, item 24. On item 24 the MW Black group did more poorly. In other cases, notably item 22, the absolute difference between Chicanos and others was quite small. But the net effect is clearly that their level of basic mathematical literacy, as defined in this way, is lowest of the groups studied.

White group typically showed a high order of success on these component tasks, while the two Black groups were somewhat more intermediate. There tended to be correlations between the groups in their level of success on the items, so that a task which was more difficult for one group was also more difficult for the others.

The topics of inequalities and of Geometry seemed to be most difficult for all groups; the set of Miscellaneous questions was also difficult. It is not easy to see in the patterns of these data any special group-content interactions. By and large, the content area summaries, the subset means reported in the parentheses below the data for a question-set, show values which rank in difficulty in highly similar ways for the groups.

Appendix A presents the distributions of the most frequently observed responses for each free response question in the quantitative measure. In each case, the data for first (correct) response reports the same percentages as those in Table 19. The subsequent categories list major specific wrong answers, together with the frequency of response in proportions. These specific responses were selected either because a lot of respondents made them or because of some presumed logical nexus to the problem. Wrong answers not

Table 21

Percentages of Correct Responses to Quantitative Items in the Supplementary Measure

	Ad hoc subject group				
Test Item	White (N=98)	Chicano (N=92)	S. East Black (N=94)	M. West Black (N=38)	Chance percent level
		Parenthetical Notation			
1. 3 (5-6) = ____	94	52	65	84	0
2. 3 (4) - 7 = ____	96	76	89	92	0
3. 4 (6-6) = ____	97	70	88	87	0
4. 5 (2) - 2 = ____	99	76	97	95	0
5. -5 (-3) = ____	98	43	81	79	0
(Subset mean)	(96.8)	(63.4)	(84.0)	(87.4)	0
		Fractions			
6. 1/2 x 4/7 = ____	88	43	79	66	0
7. $\frac{3 \times 4 \times 5}{2 \times 3}$ = ____	94	79	95	92	0
(Subset mean)	(91.0)	(61.0)	(87.0)	(79.0)	0
		Roots and Powers			
8. $\sqrt{9}$ = ____	100	74	95	97	0
9. $\sqrt{25}$ = ____	100	75	97	100	0
10. $\sqrt{1}$ = ____	93	73	89	92	0
11. 2^3 = ____	100	67	90	89	0
12. 3^2 = ____	89	66	89	89	0
13. 1^4 = ____	91	63	83	89	0
(Subset mean)	(95.5)	(69.7)	(90.5)	(92.7)	0
		Elementary Algebra			
14. If $\frac{N}{20}$ = 100, N = ____	91	48	68	66	0
15. What is the sum of y and y + z? ____	80	41	67	71	0
16. If x = -2, then x + 5 = ____	97	64	87	84	0
(Subset mean)	(89.3)	(51.0)	(74.0)	(73.7)	0

Table 21 (cont.)

Percentages of Correct Responses to Quantitative Items in the Supplementary Measure (Continued)

	Ad hoc subject group				
Test Item	White (N=98)	Chicano (N=92)	S. East Black (N=94)	M. West Black (N=38)	Chance percent level
		Averages			
17. The average of 1, 3, and 5 is _____	95	61	85	74	0
18. The average of 1, 2, 3, 10, and 14 is _____	89	45	77	71	0
19. The average of 1, 2, 4, and 5 is _____	96	53	83	82	0
(Subset mean)	(93.3)	(53.0)	(81.7)	(75.7)	
		Geometry			
20. How many degrees are there in a right angle? _____	90	52	82	82	0
21. What is the total number of degrees of arc in a circle? _____	80	43	52	55	0
(Subset mean)	(85.0)	(47.5)	(67.0)	(68.5)	
		Inequalities[a]			
27. If $10 < x < 20$, ...	89	38	63	74	0.2
28. If $-2 \leq y$, ...	70	25	32	32	< 0.1
29. If $z < -3$, ...	79	36	62	68	< 0.1
(Subset mean)	(79.3)	(33.0)	(52.3)	(58.0)	
		Miscellaneous, in Multiple-Choice form[a]			
22. (Identify consecutive integers)	92	59	62	63	17
23. (Understand $\parallel$ symbol)	92	58	73	76	25
24. Identify isosoles triangle)	64	37	43	26	20
25. (Identify specified coordinate point)	76	46	63	68	20
26. (Recognize appropriate ratio)	96	60	72	79	17
(Subset mean)	(84.0)	(52.0)	(62.6)	(62.4)	(19.7)

[a] See Appendix A for complete items

selected in this way are grouped as "miscellaneous." Blanks (omits) are also reported.

Observed frequencies for wrong answers are affected by a number of factors. One such factor is clearly the difficulty of the item. In comparing the proportion of respondents who exhibit any particular wrong answer, then, these raw proportions will need to be evaluated in the context of associated differences in number right.

A straightforward way to approach the problem is to compute the proportion demonstrating a given wrong answer, basing the computation on only the group who gave nonright answers. This gives an index of relative strength of a given response for one ethnic group. Thus, in Item 1 in Appendix A the wrong answer +3 was given by 28.7 percent of S.E. Blacks and 20.2 percent of Chicanos. But this wrong answer constituted 8[illegible] percent of all wrong answers for S.E. Blacks, versus 45.2 percent of nonright answers for Chicanos. The second contrast, then, heightens the impression that the groups differ in this tendency. To facilitate such comparisons, Table 22 was developed. This lists any wrong answers to items which showed a group difference of 20 percent or more in the index "proportion of all wrong responses consisting of this response." The analysis was arbitrarily restricted to groups where 10 or more respondents showed specific wrong answers.

The differences in Table 22 tend to contrast the Chicano and S.E. Blacks, as the least able groups, and tend to support the view that S.E. Black errors, more so than Chicano errors, demonstrate an "overt" faulty process. That is, more of the S.E. Black errors can be linked to plausible but faulty operations, such as 3(4)-7 = 0, in Item 2, where 3(4) is misprocessed as (3+4)-7 = 0, or such as failing to divide to get the average in Item 17, so that 1+3+5 = 9 is computed and 9 is made the answer.

Beyond these contrasting tendencies for the group of S.E. Blacks, the distributions show few dramatic differences. Individual findings are of some interest in assessing the needs of groups for specific counsel. On Items 10, 11, and 12, for example, while none of these Whites encountered any problems with 23 and 32, 9 Whites made the error 14 = 4. Clearly a special literacy in "powers of 1" cannot be presumed from a more general mastery of exponents.

Table 22

Specific Responses Showing Differences Larger

Than 20% in Relative Attractiveness

Item	Correct Response	Incorrect Response	Ad hoc Group	Proportion*
1	-3	3	S.E. Black Chicano	81.8 45.2
2	5	0	S.E. Black Chicano	40.0 9.1
9	1, ±1,	0	S.E. Black Chicano	10.0 32.0
14	2,000	5	S.E. Black M.W. Black	40.0 15.4
17	3	9	S.E. Black Chicano M.W. Black	7.1 11.1 20.0
17	3	4.5	Chicano M.W. Black	8.3 30.0
18	6	30	S.E. Black Chicano	27.3 3.9
19	3	12	S.E. Black Chicano	43.8 4.7
20	90, 90%	180	S.E. Black Chicano White	35.8 6.8 60.0
21	360, 360%	180	S.E. Black Chicano M.W. Black White	60.0 25.0 58.8 75.0
28	[-2, -1, 0, 1, 2, 4]	[-4, -2]	Chicano White	5.8 35.5

*Proportion of all errors consisting of incorrect response.

The linking of these results to differences in item performance would be difficult. The knowledges assessed here are basically components of the more complex GRE items. But the response distributions of the S.E. Blacks and the Chicanos indicate that the former group may be much more susceptible to certain kinds of distracters, since they appear to generate errors based on more overt process problems. Chicanos might demonstrate a flatter response pattern. In terms of the analysis variables of Part I, Chicano response distributions would be expected to show higher values of R.U. In actual fact, the Black sample in Phase I showed slightly higher R.U. values than the Chicano group. But Phase I consisted of a national sample, and the comparability of the similarly-named groups in the two samples is dubious. It cannot be established whether there are qualitative differences in the way these ethnic groups generate wrong answers to quantitative items.

Summary and Conclusions

This study was basically a search for evidence of intergroup differences in responses to tests and to test-like tasks. Essentially, it was a composite of seven component inquiries. Thus, there were three substudies in Phase I and four in Phase II. These seven studies cover the GRE item-type domain from a number of viewpoints relevant to implicit guessing beh[illegible] studies in Phase I centered on item analytic strateg[illegible] test data derived from an actual GRE administration. Th[illegible]es in Phase II centered on "item-component" strategies and o[illegible] derived from special administrations.

The common theme of all of these inquiries was the quest for indications of intergroup differences. The groups considered were Whites, Chicanos, and Blacks. It should be noted, however, that the groups studied in Phase II were ad hoc samples with no real potential for generalizability to the total populations. The implication of any differences might be that the scoring formula and the instructions to candidates concerning scoring might be inappropriate for one or more groups. The most general conclusion of the study is simply that such intergroup differences do not exist. In seven attempts to find group contrasts, with each attempt yielding a fairly complex and multifaceted analysis, only one minor phenomenon can be reported: Chicano female omitting on GRE-V is demonstrated by groups of somewhat lower ability than those demonstrating similar behavior for other groups.

These findings must be considered reassuring in the context of the continual concern that tests are less appropriate for minority students because they are approached in a different manner by these students. There is no real evidence that the minority groups here studied were different from their majority counterparts in any way not explained by overall test performance itself. Minorities score less well. But minorities do not show test performances, in doing so, which are different in process from low-scoring majority groups.

The findings are in a sense disappointing in that they do not indicate that test differences can be reduced by adjustments of the scoring and instructional techniques. It is widely believed that test score differences on measures like the GRE are not true differences in developed education potential, and there is a quest for some adjustment in the testing situation which will reduce the score difference. This study fails to demonstrate the sought-for change. On the other hand, its results support the view that the present test configuration may be in fact less biased than its critics suggest, since in seven substudies, whose detailed and systematic comparisons offered a large number of potential opportunities for demonstrable differences, only two limited "quasi-successes" were registered.

While the results are of sufficient interest to suggest follow-up studies, they cannot explain score differences in any practical way. In evaluating the seven indices used in these studies, it must be recognized that each is a rational window on candidate behavior but that there is no prior empirical work validating them, upon which the study could build. The rationales presented in the report contribute to the face validity of the measures, but the limitations of the instruments are clear. Such seemingly simple steps as the "recombining" of analogy item associational data by averaging "association" strength across the two members of a response-pair are themselves in need of research verification.

In spite of the limitations of the indicators, however, there is some evidence that they are meaningfully related to the internal item process variables of candidates. In the Phase I analysis, this was most strongly indicated by the item format effects. The P+R and P+A variables are widely used in testing, and item format differences for them have been long established. But the MnO and the RU indicators are not so well established, and the finding of format effects substantiates the view that these indicators are sensitive to process shifts. In Phase II, the limited test of the link between association ratings and item performance in analogies attests to the ability of the association approach to reflect internal process. The rough associational measure seems to have promise of value in explaining the popularity of distractor choices; the "mathematical literacy" measure seemed to reflect some group differences in free response data which are process-oriented. On the whole, the measures developed for this study seemed to "work" in the sense of yielding differentiation and a potential thereby for reflecting group characteristics.

While the principle findings are reassuring concerning the current structure of the test, the study should serve as an incentive to further work. The impetus here was toward an examination of intergroup differences in item response process. The focus was on guessing process, and the tone and tenor of the study was on the evaluation of the existing program practices in instruction and in scoring. But item process investigations have a valid role of their own. There is little clear knowledge of why aptitude measures work, of how they are solved. Work by Bloom and Broder (1950) and more recently by Sternberg (1977) is related to the current study, as is the extensive effort to infer item solution process from item analysis by Brigham (1930) in his book "The Study of Error." The work in the present study covers a number of facets of the GRE item domain. But this very breadth necessarily leads to constraints in the number of any given item type which can be surveyed. The study of truncated sentence completion items, for example, had to be based on a very few of these items. The finding that the different groups gave virtually equivalent evaluations of distractor potential on this task is an intriguing one in view of the known group differences in success on the items themselves. Additional work with item component studies and sentence completions could expand the item material (reduce the truncation) to see if there is some critical information level which triggers process differences. As matters stand, an inferential task seemingly as hard as the item task itself is found to demonstrate virtually no group differences, while items themselves show such differences.

In summary, the study suggests that

(1) there are no significant differences in the response processes of ethnic groups as these relate, implicitly, to guessing;

(2) there is one minor group-linked difference for Chicano females on Verbal material, perhaps worthy of some further study;

(3) there is a genuine potential for understanding item solution processes, not merely those related to guessing, through item component studies of the type embodied in the special measures of Phase II. Additional work with such measures is justified and desirable.

References

Bloom, B., & Broder, L. Problem-solving processes of college students. Chicago: University of Chicago Press, 1950.

Brigham, C. C. A study of error. New York: College Entrance Examination Board, 1932.

Campbell, J. T., & Belcher, L. H. Word associations of students at predominantly White and predominantly Black colleges. Research Bulletin 75-29. Princeton, N. J.: Educational Testing Service, 1975.

Conrad, L. & Wal[illegible], M. M. Report to the Research Committee on the item analysis of a GRE Aptitude Test by ethnic and sex subgroups. GRE Board Research Committee materials for September, 1975. (mimeo)

Cureton, E. E. Reliability of multiple-choice tests is the proportion of variance which is true variance. Educational and Psychological Measurement, 1971, 31, 827-829.

Echternacht, G. J., Carlson, A. B., & Flaugher, R. L. An inconclusive study of differences in test and item performance for Black and White undergraduates. Unpublished, 1972.

Flaugher, R. L., & Pike, L. W. Reactions to a very difficult test by an inner-city high school population: A test and item analysis. Research Memorandum 70-11. Princeton, N. J.: Educational Testing Service, 1970.

Moore, J. C. Test-wiseness and analogy test performance. Measurement and Evaluation in Guidance, 1971, 3, 198-202.

Peterson, R. E. Technical manual: College Student Questionnaires. Princeton, N. J.: Educational Testing Service, 1965.

Pike, L. W., & Flaugher, R. L. Assessing the meaningfulness of group responses to multiple-choice test items. Proceedings, 78th Annual Convention, American Psychological Association, 1970.

Shannon, C. E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27, 379-423, 623-656.

Slakter, M. J., Crehan, K. D., & Koehler, R. A. Longitudinal studies of risk taking on objective examinations. Educational and Psychological Measurement, 1975, 35, 97-105.

Sternberg, R. J. Component processes in analogical reasoning. *Psychological Review*, 1977, *84*, 353-378.

Swineford, F., & Miller, P. M. Effects of directions regarding guessing on item statistics of a multiple-choice vocabulary test. *Journal of Educational Psychology*, 1953, *44*, 129-139.

Willner, A. An experimental analysis of analogical reasoning. *Psychological Reports*, 1964, *15*, 479-494.

Appendix A

Percentages of Responses to Quantitative Items, by Experimental Group and by Individual Response Category

Response[1]	Experimental Group SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 1:3 (5-6) =				
-3*	65	52	84	94	72
3	29	[illegible]	8	4	16
9	1	5	0	0	2
4	0	4	0	0	1
Misc.	4	12	8	2	6
Omit	1	4	0	0	2
	Item 2:3 (4)-7 =				
5*	89	76	92	96	88
0	4	2	0	0	2
4	2	1	3	0	1
-9	0	0	0	1	0
Misc.	4	14	5	34	7
Omit	0	6	0	0	2
	Item 3:4 (6-6) =				
0*	88	70	87	97	85
4	6	20	8	3	9
18	2	3	3	0	2
Misc.	3	1	3	0	2
Omit	0	7	0	0	2

[1] Correct responses are marked with an asterisk.

Note: The answer choices have been reordered most popular to least popular.

Experimental Group

Response	SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 4:5 (2) -2 =				
8*	97	76	95	99	91
-20	1	3	0	1	2
0	1	1	3	0	1
Misc.	1	12	3	0	4
Omit	0	8	0	0	2
	Item 5: 1/2 x 4/7 =				
2/7 or 4/14*	79	44	66	88	70
15/14	4	8	3	3	5
8/7	2	1	0	1	1
7/8	2	2	0	0	1
5/14	1	1	[illegible]	0	1
5/9	1	0	3	0	1
Misc.	8	23	13	4	11
Omit	3	22	13	4	10
	Item 6: -5 (-3)				
15 or +15*	81	44	79	88	72
-15	10	25	11	0	11
-2	3	6	0	3	4
2	1	[illegible]	0	0	1
Misc.	3	5	5	5	5
Omit	2	14	5	4	7

		Experimental Groups			
Response	SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 7: $\sqrt{9}$ =				
3 or ± 3*	95	74	97	100	91
Misc.	4	14	3	0	5
Omit	1	12	0	0	4
	Item 8: $\sqrt{25}$ =				
5 or ± 5*	97	75	100	100	92
Misc.	2	13	0	0	4
Omit	1	12	0	0	4
	Item 9: $\sqrt{1}$ =				
1 or ± 1*	89	73	92	93	86
0	1	9	0	1	3
1/2	0	0	0	2	1
Misc.	1	2	5	3	2
Omit	9	16	3	1	8
	Item 10: 2^3 =				
8*	90	67	90	100	87
6	4	9	0	0	4
Misc.	3	12	10	0	6
Omit	2	12	0	0	4
	Item 11: 3^2 =				
9*	89	66	90	89	83
27	4	4	5	11	7
6	3	9	0	0	3
Misc.	1	9	5	0	3
Omit	2	12	0	0	4

Experimental Group

Response	SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 12: 1^4 =				
1*	83	63	89	91	80
4	15	22	5	9	14
Misc.	0	5	3	0	2
Omit	2	10	3	0	4
	Item 13: $\frac{3 \times 4 \times 5}{2 \times 3}$ =				
10 or 60/6 or 30/3*	95	79	92	94	90
12	2	0	0	2	1
Misc.	3	11	5	4	6
Omit	0	10	3	0	3
	Item 14: If $\frac{N}{20}$ = 100, N =				
2000*	68	48	66	91	69
5	13	13	5	2	9
200	3	2	21	4	5
1000	3	3	3	0	2
20,000	2	0	0	2	1
Misc.	2	5	0	0	2
Omit	9	28	5	1	12
	Item 15: What is the sum of y and y + z?				
2y + z or 2y + 3*	67	41	71	80	64
y + y or y + (y + z)	5	3	5	3	4
y^2 + 2	4	3	0	2	3
y2	3	5	0	0	2
y + z	3	1	0	0	1
Misc.	7	17	11	10	12
Omit	10	28	13	5	14

ERIC Full Text Provided by ERIC

	Experimental Groups				
Response	SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 16: If x = -2, then x + 5 =				
3*	87	64	84	97	83
-3	3	2	3	1	2
7	3	1	3	1	2
Misc.	4	12	3	1	5
Omit	2	21	8	0	8
	Item 17: The average of 1, 3 and 5 is				
3*	85	61	74	95	80
9	9	4	5	0	4
4.5	3	3	8	4	4
2	1	1	0	0	1
Misc.	1	13	8	1	5
Omit	1	17	5	0	6
	Item 18: The average of 1, 2, 3, 10 and 14 is				
6*	77	45	71	89	70
30	6	2	5	0	3
15	3	2	5	3	3
5	4	1	8	0	3
8	1	1	3	3	2
3	1	3	0	0	1
4	2	1	3	0	1
Misc.	3	17	0	5	8
Omit	2	27	5	0	9

	Experimental Group				
Response	SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 19: The average of 1, 2, 4, and 5 is				
3*	83	53	82	96	78
12	7	2	5	0	3
6	3	3	5	3	3
Misc.	3	19	5	1	7
Omit	3	23	3	0	8
	Item 20: How many degrees are there in a right angle?				
90,* 90%	82	52	82	90	76
180	6	3	10	6	6
45	2	11	3	3	5
60	2	1	0	0	1
30	1	1	0	0	1
Misc.	5	11	3	1	5
Omit	1	21	3	0	6
	Item 21: What is the total number of degrees of arc in a circle?				
360,* 360%	52	44	55	80	58
180	29	14	26	15	20
45	2	2	0	0	1
0	2	1	0	0	1
120	1	1	0	0	1
Misc.	4	13	10	4	8
Omit	10	25	8	1	11

ERIC

Response	Experimental Group SE Black (N = 94)	Chicano (N = 92)	MW Black (N = 38)	White (N = 98)	All Ss (N = 322)
	Item 27:	If 106 x < 20, which of the following can be the values of x? 5, 10, 15, 20, 25, 30			
15*	63	38	74	89	65
5	11	4	10	3	6
10	1	6	[illegible]	0	2
5, 10, 15	2	2	0	2	2
Misc.	7	9	8	3	6
Omit	16	40	5	3	18
	Item 28:	If -2 ≤ y, which of the following can be values of y? -4, -2, -1, 0, 1. 2, 4			
-2, -1, 0, 1, 2, 4*	32	25	32	70	42
-4, -2	17	4	16	10	11
-2	11	3	16	1	6
-1	4	2	8	1	3
-4	2	5	5	0	3
-1, 0, 1, 2, 4	3	2	3	3	3
1	2	1	0	0	1
Misc.	7	9	8	6	8
Omit	21	48	13	8	24
	Item 29:	If z < -3, which of the following can be values of z? -6, -3, -1, 0, 1, 3, 6			
-6*	62	36	68	79	60
-3	3	4	3	0	2
-1	5	0	8	0	2
-6, -3	0	3	3	2	2
-1, 0, 1, 3, 6	0	2	3	3	2
-6, -3, -1	2	0	0	1	1
Misc.	4	6	3	4	5
Omit	23	48	13	11	26

Appendix B

Instrument Used to Collect Supplementary Data (Phase II)

NAME________________________ SCHOOL________________________

DATE________________________

BACKGROUND INFORMATION

In order to make the best use of the experimental data, we will need to have the background information requested in the next four questions. You may, however, leave any questions unanswered that you wish.

Mark your answers by circling the appropriate number for each question.

A. Please indicate your sex

1. Male
2. Female

B. What is your present educational status?

1. Freshman
2. Sophomore
3. Junior
4. Senior
5. Other (Specify)________________

C Which best describes your major field of study?

1. Humanities (Art, English, languages, philosophy, etc.)
2. Social Sciences (Education, history, government, law, psychology, etc.)
3. Biological Sciences (Agriculture, biology, forestry, home economics, nursing, etc.)
4. Physical Sciences (Mathematics, physics, chemistry, engineering, computer sciences, etc.)

D. When did you last attend high school on at least a half-time basis?

1. One to three years ago
2. Four to six years ago
3. Seven to nine years ago
4. Ten or more years ago

WORD ASSOCIATIONS MEASURE

Directions. For each word given in capital letters, you are to judge which of the five following words is most closely associated or related to it, and indicate this by circling the 5 in the appropriate row. Similarly, you are to circle the 4 for the word that is next-most related, and so on. Look at the example:

Example

RICH		Least Related				Most Related
a.	calendar	[✓ 1]	2	3	4	5
b.	cucumber	1	2	[✓ 3]	4	5
c.	poor	1	2	3	4	[✓ 5]
d.	sound	1	[✓ 2]	3	4	5
e.	wealthy	1	2	3	[✓ 4]	5

The words "poor" and "wealthy" would come rather quickly to mind for most of us, when given the word "rich." Those feeling that "poor" is the most closely related would circle the 5 for that word. Others might reverse the order between "poor" and "wealthy." They would circle the 5 for "wealthy" and the 4 for "poor." There are likely to be a lot of individual differences in marking the other three relationships. "Cucumber" can be related in the sense of rich food, "sound" in the sense of high fidelity, and "calendar" in the sense that, for example, the "rich" may be though as being a slave to clock and calendar.

When you are rating the level of relatedness of each set of words, please keep these points in mind:

1. There are no right or wrong answers.

2. Circle one number for each word. If you don't know the meaning of a word or are otherwise unsure, follow your "hunch" about the level of relatedness.

3. For each set of five words, use each level of relatedness only once. "Toss a coin" when there is a tie.

4. Remember that 1 indicates least related, and that 5 indicates most related.

WORD ASSOCIATIONS MEASURE

A. SONG

		Least Related				Most Related
1.	score	1	2	3	4	5
2.	instrument	1	2	3	4	5
3.	solo	1	2	3	4	5
4.	benediction	1	2	3	4	5
5.	suit	1	2	3	4	5

B. SKUNK

		Least Related				Most Related
6.	camel	1	2	3	4	5
7.	porcupine	1	2	3	4	5
8.	lion	1	2	3	4	5
9.	cat	1	2	3	4	5
10.	hound	1	2	3	4	5

C. WATER

11.	cloud	1	2	3	4	5
12.	power	1	2	3	4	5
13.	matter	1	2	3	4	5
14.	antiques	1	2	3	4	5
15.	valuables	1	2	3	4	5

D. IDEA

16.	thought	1	2	3	4	5
17.	play	1	2	3	4	5
18.	theory	1	2	3	4	5
19.	dream	1	2	3	4	5
20.	attention	1	2	3	4	5

E. HISTORIAN

21.	skeptic	1	2	3	4	5
22.	logician	1	2	3	4	5
23.	politician	1	2	3	4	5
24.	archaeologist	1	2	3	4	5
25.	statistician	1	2	3	4	5

F. OBSESSION

26.	emotion	1	2	3	4	5
27.	author	1	2	3	4	5
28.	experimentation	1	2	3	4	5
29.	thought	1	2	3	4	5
30.	vigil	1	2	3	4	5

G. CONFIDENCE

31.	simplicity	1	2	3	4	5
32.	economy	1	2	3	4	5
33.	conscience	1	2	3	4	5
34.	fear	1	2	3	4	5
35.	peace	1	2	3	4	5

H. SCENT

36.	hump	1	2	3	4	5
37.	quill	1	2	3	4	5
38.	mane	1	2	3	4	5
39.	whisker	1	2	3	4	5
40.	ear	1	2	3	4	5

Please go on to the next page.

WORD ASSOCIATIONS MEASURE

I. SUPPLICATE	Least Related				Most Related
41. ingratiate	1	2	3	4	5
42. request	1	2	3	4	5
43. demand	1	2	3	4	5
44. peruse	1	2	3	4	5
45. entreat	1	2	3	4	5

J. REPERTOIRE	Least Related				Most Related
46. melody	1	2	3	4	5
47. artist	1	2	3	4	5
48. chorus	1	2	3	4	5
49. church	1	2	3	4	5
50. wardrobe	1	2	3	4	5

K. CISTERN					
51. shower	1	2	3	4	5
52. official	1	2	3	4	5
53. science	1	2	3	4	5
54. museum	1	2	3	4	5
55. vault	1	2	3	4	5

L. HUMBLE					
56. thankful	1	2	3	4	5
57. solution	1	2	3	4	5
58. peremptory	1	2	3	4	5
59. cursory	1	2	3	4	5
60. aggressive	1	2	3	4	5

M. ODIUM					
61. amusement	1	2	3	4	5
62. loathing	1	2	3	4	5
63. mortification	1	2	3	4	5
64. ingratitude	1	2	3	4	5
65. envy	1	2	3	4	5

N. ANACHRONISM					
66. atheism	1	2	3	4	5
67. fallacy	1	2	3	4	5
68. propaganda	1	2	3	4	5
69. artifact	1	2	3	4	5
70. criterion	1	2	3	4	5

O. DIFFIDENT					
71. affluent	1	2	3	4	5
72. profligate	1	2	3	4	5
73. cunning	1	2	3	4	5
74. antagonistic	1	2	3	4	5
75. moribund	1	2	3	4	5

P. INVIDIOUS					
76. droll	1	2	3	4	5
77. winsome	1	2	3	4	5
78. pious	1	2	3	4	5
79. benign	1	2	3	4	5
80. unpretentious	1	2	3	4	5

STOP. Make sure you have circled a number for each word, and that for each set of 5 words, you have used all 5 levels of relatedness.

5

USING CONTEXTUAL CLUES

Directions. In each of the next questions, you are given part of a sentence, with a blank where a word has been removed, and a list of five words. You are to circle the 5 for the word that is most likely to belong in the blank, the 4 for the word that is next-most likely, and so on.

Be sure to circle a different number for each word. "Toss a coin" when there is a tie.

Remember that 1 indicates the least likely substitution, and that 5 indicates the most likely substitution.

Example:

. . . accelerate the ____________ of soil nutrients . . .

	Least likely				Most likely
A. depletion	1	2	3	4	[✓ 5]
B. erosion	1	2	[✓ 3]	4	5
C. cultivation	1	[✓ 2]	3	4	5
D. fertilization	[✓ 1]	2	3	4	5
E. conservation	1	2	3	[✓ 4]	5

A. . . . growth is not a ____________ process for all people . . .

1. uniform	1	2	3	4	5
2. healthy	1	2	3	4	5
3. unique	1	2	3	4	5
4. simple	1	2	3	4	5
5. progressive	1	2	3	4	5

B. . . . problems that are difficult today may become ____________ tomorrow . . .

6. insoluble	1	2	3	4	5
7. manageable	1	2	3	4	5
8. dissipated	1	2	3	4	5
9. prominent	1	2	3	4	5
10. vital	1	2	3	4	5

Please go on to the next page.

Using Contextual Clues

C. . . . pervasive feature of human intellect is its ______________capacity. . .

		Least likely				Most likely
11.	finite	1	2	3	4	5
12.	inadequate	1	2	3	4	5
13.	remarkable	1	2	3	4	5
14.	boundless	-	2	3	4	5
15.	limited	1	2	3	4	5

D. . . . geese________________ no territorial behavior.

16.	identify	1	2	3	4	5
17.	merit	1	2	3	4	5
18.	divulge	1	2	3	4	5
19.	accept	1	2	3	4	5
20.	manifest	1	2	3	4	5

E. . . . he was famous for his ________________and irresponsible behavior . . .

21.	bizarre	1	2	3	4	5
22.	penurious	1	2	3	4	5
23.	cavalier	1	2	3	4	5
24.	obsequious	1	2	3	4	5
25.	licentious	1	2	3	4	5

STOP. Make sure you have circled a number for each word, and that for each set of 5 words, you have used all 5 levels of likelihood.

ERIC Full Text Provided by ERIC

RECOGNITION VOCABULARY

Directions. Rate each word given below on the seven-point scale that is provided. A rating of 1 indicates words that you least recognize. It should be used for words with which you are completely unfamiliar. A rating of 7 indicates words with which you are very familiar, and have no doubt regarding their meaning. The middle rating, 4 , indicates words that you would probably understand in context, but are doubtful when the word appears alone.

Be sure to circle a number for each word.

Remember that 1 indicates words you least recognize, and 7 indicates words you most readily recognize.

Use all seven values, to indicate different degrees of word knowledge.

		Least Recognition						Most Recognition
1.	abrasiveness	1	2	3	4	5	6	7
2.	accomplish	1	2	3	4	5	6	7
3.	acridness	1	2	3	4	5	6	7
4.	alleviate	1	2	3	4	5	6	7
5.	anachronism	1	2	3	4	5	6	7
6.	analyze	1	2	3	4	5	6	7
7.	anarchy	1	2	3	4	5	6	7
8.	antecedent	1	2	3	4	5	6	7
9.	apprehend	1	2	3	4	5	6	7
10.	cistern	1	2	3	4	5	6	7
11.	clarify	1	2	3	4	5	6	7
12.	conciliate	1	2	3	4	5	6	7
13.	confidence	1	2	3	4	5	6	7
14.	confuse	1	2	3	4	5	6	7
15.	corroborate	1	2	3	4	5	6	7

Please go on to the next page.

Recognition Vocabulary

		Least						Most
16.	demote	1	2	3	4	5	6	7
17.	diffident	1	2	3	4	5	6	7
18.	discourage	1	2	3	4	5	6	7
19.	dispute	1	2	3	4	5	6	7
20.	distort	1	2	3	4	5	6	7
21.	doubt	1	2	3	4	5	6	7
22.	ecclesiastical	1	2	3	4	5	6	7
23.	estrange	1	2	3	4	5	6	7
24.	exigencies	1	2	3	4	5	6	7
25.	expedience	1	2	3	4	5	6	7
26.	fluidity	1	2	3	4	5	6	7
27.	habitual	1	2	3	4	5	6	7
28.	heinous	1	2	3	4	5	6	7
29.	hiatuses	1	2	3	4	5	6	7
30.	historian	1	2	3	4	5	6	7
31.	hominid	1	2	3	4	5	6	7
32.	humble	1	2	3	4	5	6	7
33.	hyperboles	1	2	3	4	5	6	7
34.	idea	1	2	3	4	5	6	7
35.	inaugurations	1	2	3	4	5	6	7
36.	incommodities	1	2	3	4	5	6	7
37.	incomprehensible	1	2	3	4	5	6	7
38.	implement	1	2	3	4	5	6	7
39.	intensify	1	2	3	4	5	6	7
40.	invidious	1	2	3	4	5	6	7
41.	involuntary	1	2	3	4	5	6	7
42.	negotiate	1	2	3	4	5	6	7
43.	obsession	1	2	3	4	5	6	7
44.	odium	1	2	3	4	5	6	7
45.	pardonable	1	2	3	4	5	6	7

Please go on to the next page.

Recognition Vocabulary

	Least						Most
46. paucity	1	2	3	4	5	6	7
47. placate	1	2	3	4	5	6	7
48. plenum	1	2	3	4	5	6	7
49. profanity	1	2	3	4	5	6	7
50. progenitorship	1	2	3	4	5	6	7
51. proliferation	1	2	3	4	5	6	7
52. promulgate	1	2	3	4	5	6	7
53. reciprocate	1	2	3	4	5	6	7
54. repertoire	1	2	3	4	5	6	7
55. restore	1	2	3	4	5	6	7
56. retard	1	2	3	4	5	6	7
57. scent	1	2	3	4	5	6	7
58. secular	1	2	3	4	5	6	7
59. sell	1	2	3	4	5	6	7
60. skunk	1	2	3	4	5	6	7
61. song	1	2	3	4	5	6	7
62. supplicate	1	2	3	4	5	6	7
63. suppress	1	2	3	4	5	6	7
64. surfeit	1	2	3	4	5	6	7
65. temporality	1	2	3	4	5	6	7
66. tenacity	1	2	3	4	5	6	7
67. uncontrollable	1	2	3	4	5	6	7
68. vacuum	1	2	3	4	5	6	7
69. water	1	2	3	4	5	6	7

STOP. Make sure you have circled a number for each word.

QUANTITATIVE MEASURE

A. Answer the following questions, by filling in the blanks.

1. 3(5-6) = ________
2. 3(4)-7 = ________
3. 4(6-6) = ________
4. 5(2)-2 = ________

5. 1/2 x 4/7 = ________
6. -5(-3) = ________
7. $\sqrt{9}$ = ________
8. $\sqrt{25}$ = ________

9. $\sqrt{1}$ = ________
10. 2^3 = ________
11. 3^2 = ________
12. 1^4 = ________

13. $\frac{3 \times 4 \times 5}{2 \times 3}$ = ________

14. If $\frac{N}{20}$ = 100, N = ________

15. What is the sum of y and y + z? ________
16. If x = -2, then x + 5 = ________
17. The average of 1, 3, and 5 is ________
18. The average of 1, 2, 3, 10, and 14 is ________
19. The average of 1, 2, 4, and 5 is ________
20. How many degrees are there in a right angle?________
21. What is the total number of degrees of arc in a circle? ________

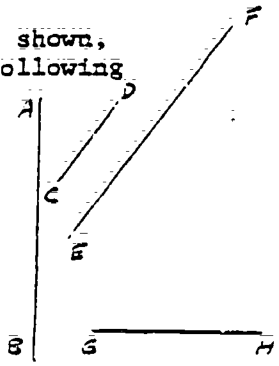
B. Answer the following questions by circling the one correct answer to each.

22. Which of the following is a set of consecutive integers?

1. 1.1 and 1.2
2. 1.2 and 1.4
3. 1.3 and 1.5
4. 11 and 12
5. 12 and 14
6. 13 and 15

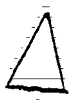
23. For the figure shown, which of the following is true?

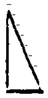
1. AB ‖ CD
2. AB ‖ EF
3. CD ‖ EF
4. EF ‖ GH

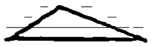


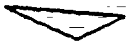
24. Which of the following is an isosoles triangle?

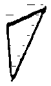
1. A
2. B
3. C
4. D
5. E


A


B


C

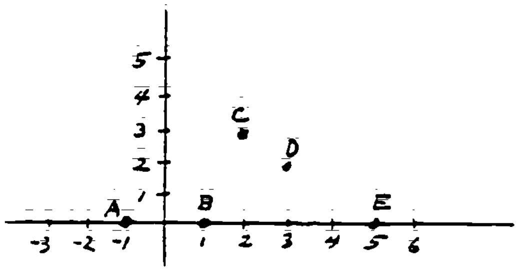

D


E

Please go on to the next page.

25. In the coordinate system shown, which point has the coordinates 2, 3?

1. A
2. B
3. C
4. D
5. E



26. If there are 3 men, 2 women, and 1 child in a group, what is the ratio of men to women?

1.	2/3	4.	3/2
2.	5/6	5.	5
3.	1	6.	6

C. Answer the following questions by circling all the correct answers for each.

27. If 10 < x < 20, which of the following can be values of x?

5, 10, 15, 20, 25, 30

28. If -2 ≤ y, which of the following can be values of y?

-4, -2, -1, 0, 1, 2, 4

29. If z <-3, which of the following can be values of z?

-6, -3, -1, 0, 1, 3, 6

STOP. Check your answers to the quantitative questions. Then close your test booklet and await further instructions.

Appendix B: Judges' instructions for rating analogy choices on their shared relationships with the item pair.

In each of the following questions, a related pair of words of phrases is followed by five lettered pairs of words or phrases. Select the lettered pair which best expresses a relationship similar to that expressed in the original pair, and mark a rating of 5 for that choice. Then select the lettered pair which is next best in expressing a relationship similar to that expressed in the original pair, and mark it with a 4, and so on. Do not allow ties. If two choices are very similar in your judgment, give them different ratings, but note marginally that the two were very close.

To facilitate subse‹.ıent discussion and resolution of differences in ratings, it would be helpful to have brief relational statements indicating the educed relationship for each choice pair that was compared to relationships in the stem pair.

EXAMPLE	Rating (5=best)	Relational statements; comments
REQUEST: ENTREAT ::		
(A) control : explode	2	control - loosely, the opp. of - explode
(B) admire : idolize	5	admire - generic of the intensive - idolize
(C) borrow : steal	4	borrow - legal variant of the illegal ~~act~~ - steal
(D) repeat : plead	1	repeat - {~~no evident relationship to~~ (x "goes with" y ?) } - plead
(E) cancel : invalidate	3	cancel - a special type/case of - invalidate

note - 1&2 were very close