DOCUMENT RESUME

ED 388 722 TM 024 187

AUTHOR	Angoff, William H.; Johnson, Eugene G.
TITLE	A Study of the Differential Impact of Curriculum on Aptitude Test Scores.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-46
PUB DATE	Aug 88
NOTE	135p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Aptitude Tests; College Students; *Curriculum; Higher Education; Majors (Students); *Mathematics Tests; *Scores; Sex Differences; Test Results; *Verbal Tests
IDENTIFIERS	*Graduate Record Examinations; *Scholastic Aptitude Test

ABSTRACT

A sample of 22,923 students who had taken the Graduate Record Examination (GRE) General Test in the academic years 1983-84 and 1984-85, and who had also taken the Scholastic Aptitude Test (SAT) 4 or 5 years earlier was identified and classified by undergraduate field of study (four major curriculum categories) and sex. Several analyses were undertaken to determine the degree of differential impact that sex and field of study might have on GRE-verbal, GRE-quantitative, and GRE-analytical scores, after controlling on SAT-verbal and SAT-mathematical scores. It was found that correlations of SAT-verbal with GRE-verbal and SAT-mathematical with GRE-quantitative were extremely high for the entire sample and for eight identified subgroups. The impact of curriculum and sex was found to be low for GRE-verbal, but relatively high for GRE-quantitative, and moderate for GRE-analytical. Additional studies concentrating only on clearly verbal and clearly mathematical fields showed small additional impact. Another study indicated that there was a generally slight effect of the institution attended on GRE-quantitative scores, but the basic study conclusions remained unchanged. An appendix lists the major fields. (Contains 35 tables, 10 figures, and 19 references.) (SLD)

RESEARCH REPORT

A STUDY OF THE DIFFERENTIAL IMPACT OF CURRICULUM ON APTITUDE TEST SCORES

William H. Angoff
Eugene G. Johnson

Educational Testing Service
Princeton, New Jersey
August 1988

A Study of the Differential Impact of Curriculum on Aptitude Test Scores

William H. Angoff

and

Eugene G. Johnson

Educational Testing Service

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Nancy J. Robertson and John J. Ferris for their creative work in matching the scores for this huge sample, a task heretofore thought to be beyond the limits of reasonable possibility, and for their production of the statistical output that made this study possible. They also wish to express their thanks to Benjamin King for his invaluable contributions in the initial design of the analysis. Finally, the authors wish to express their appreciation to Mary L. Varone for her patience and expertise in coping with the numerous revisions required in the production of this report.

This project was supported by Educational Testing Service.

ABSTRACT

A sample of 22,923 students who had taken the GRE General Test in the academic years 1983-84 and 1984-85 and who had also taken the SAT four or five years earlier were found, and classified by undergraduate field of study (four major categories of curriculum) and sex. Several analyses were undertaken to determine the degree of differential impact that sex and field of study might have on GRE-verbal, GRE-quantitative, and GRE-analytical scores, after controlling on SAT-verbal and SAT-mathematical scores. It was found, first, that the correlations of SAT-verbal with GRE-verbal and SAT-mathematical with GRE-quantitative were extremely high, both for the entire sample, and within it, for the eight subgroups defined by field of study and sex. The correlations were .86 in the total sample and ranged from the low to middle .80s in the eight subgroups. The impact of curriculum and sex was found to be low on GRE-verbal scores, but relatively high for GRE-quantitative, with students in heavily quantitative fields enjoying an advantage over their peers in less quantitative fields of study. The impact was moderate for GRE-analytical. Further studies designed to "purify" the fields of study and include only clearly verbal fields and clearly mathematical fields--omitting entirely students in social and biological science--showed small additional impact. An additional study indicated that there was a generally slight effect of the institution attended on GRE-quantitative scores, after controlling for major field of study and initial ability, although the importance of institution attended was somewhat greater for higher ability students. Although these studies helped a bit to clarify the results, the basic conclusions remained unchanged.

In a separate phase of the study an attempt was made by means of Mantel-Haenszel analyses to identify the kinds of items that were relatively resistant to curricular and sex effects. Although the items differed from one another with respect to impact, they did not fall into identifiable categories that would make it possible to predict which items would be likely to show such impact and which would not.

INTRODUCTION

The concept of academic aptitude seems to have been invented to account for the fact that individuals who have been exposed to approximately the same educational stimuli nevertheless consistently display stable and predictable differences in achievement. Despite the fact that such differences are commonly observed, however, the concept of aptitude and its amenability to valid measurement have been subjects of considerable debate for some time, perhaps particularly in the last 15-20 years. This debate has been given new force in recent years by the appearance of additional--or to use Anastasi's (1975) word, "surplus," i.e., unwarranted and probably invalid--meanings and implications that have attached themselves to the notion of aptitude, as well as the occasionally invalid uses to which aptitude tests have sometimes been put. The implications of these surplus meanings are often articulated in popular discussions, where they have caught the attention and interest of the general public.

Opposition to the use of the concept of aptitude, even in its more conservative meanings, has often been socially and politically motivated, deriving its impetus from the commonly held view that aptitude is genetically determined. Given this view, the leap has frequently been made to assume further that aptitude is therefore unchangeable, both within a given lifetime and across generations. What has made these views objectionable politically is that they are thought to imply, one, that Blacks, for example, who typically score significantly lower than Whites in this society, are innately inferior; and two, that the low scores (and, by inference, the aptitude and intelligence) of Black parents will be followed by the low scores of their children, with the result that the intellectual and social disparities of the present will continue to be a fact of the future.

These perceptions persist, even in the face of evidence and logic to the contrary; and, curiously, the same perceptions seem to be shared by antagonistic political groups, those who are favorable to the implications and those who find them unacceptable. Unfortunately, the controversy is so charged with emotion that some potentially useful explorations into the validity (or invalidity) of the implications are often slow in coming.

Leaving the social and political issues aside for the time being, however crucial they are in other contexts, it may be useful to examine here some of the facets of the concept of aptitude that need eventually to be clarified before we can consider its usefulness as a construct in its own right. One of these has to do with its distinctiveness as a concept separate from the concept of achievement. Quite apart from this, but related to it, is the question whether it can be satisfactorily measured in a way that distinguishes it from the measurement of achievement. A second has to do with the changeability of aptitude and the nature of that changeability, either within the individual or across cohorts of individuals. Finally, a third question is the role of the genetic origins of aptitude in the matter of changeability. These are, each of them, large subjects, and no pretense is made that they will be dealt with here in exhaustive detail. At the same time, it may be helpful to examine them, however briefly.

Before doing so, it will be useful to observe again that, given the same amount of exposure to education, both inside and outside the walls of the classroom, some of us seem to be able to solve problems, understand the significance of events, facts, and connections, and draw inferences, generalizations, and deductions that others among us cannot do at all, or if they can, not so readily. It seems also to be true that although some individuals can learn the same material as others can, they do so more slowly and with more effort.

There is little question that these observations lend considerable validity to the concept of aptitude as a legitimate construct. Yet, there is a great unwillingness to accept it as such. Anastasi (1984), for example, speaks of aptitude as an "indestructible strawperson," and says that she would, if she could, excise it from our vocabulary (Anastasi, 1980). This is curious, in a sense. We seem to have no difficulty accepting other aptitudes as valid and useful constructs: athletic aptitude, musical aptitude, mechanical aptitude, artistic, and dramatic aptitude to name just a few. And just as with academic aptitude, we know that there are vast differences among us with respect to our rates of learning in these areas. Yet, while these other aptitudes are generally accepted as valid constructs, the construct of academic aptitude appears, in some quarters, at least, to be harder to accept.

In an effort to clarify the concept, some attempts have been made to develop what are thought to be clear distinctions between academic aptitude and academic achievement. For example, the College Entrance Examination Board, whose tests since its founding in 1900 had been specifically developed and used only to evaluate the student's acquired knowledge of particular secondary school subjects, introduced the Scholastic Aptitude Test in 1925.

This test was conceived as a supplement to the existing Achievement Test battery and was intended to provide a broad measure of the student's general ability to pursue <u>any</u> academic program successfully. With similar purpose the Graduate Record Examinations developed in 1952 a system of aptitude and achievement tests, the former, to measure general academic promise, and the latter, to assess what the students had learned in their particular college courses.

Nevertheless, the distinctions between aptitude and achievement are often unclear and difficult to make. It is frequently the case that constructs are easily confused with the instruments we have designed to measure them, so that we often make judgments of the validity of a construct when we are actually judging the adequacy of the instruments we use to measure it. So too here. Additionally, it is often impossible to distinguish a test of aptitude from a test of achievement; their contents are frequently so similar. Indeed, it has been observed that the tests designed to measure the concepts of aptitude and achievement are often more similar than the concepts themselves. But we do make some distinctions between both the concepts and the instruments:

1. Growth in achievement results from more-or-less formal exposure to a particular subject or area of content and is typically quite rapid. Aptitude, on the other hand, grows slowly as a consequence of ordinary living, both outside the formal learning environment as well as inside it, often developing through "unidentified and uncontrolled learning" (Anastasi, personal communication).

2. Aptitude tends to resist short-term efforts to hasten its

growth. Achievement is much more susceptible to such efforts.

3. It has often been said that scores on an achievement test are to be taken as a measure of the amount learned; aptitude tests are thought to provide a measure (or prediction) of the rate of future learning.

4. Humphreys (1974) holds, as do others, that aptitude and achievement tests differ only in degree and that specific tests of these two concepts fall on a continuum. He goes on to make essentially the following observations: Aptitude tests draw their items from a wide range of human experience. (Intelligence tests, which are a close relative of aptitude tests, draw their items from an even wider, and often different, range of experiences and include a much wider variety of items than do achievement tests.) When aptitude tests do make use of subject-matter learned in formal course work, they typically draw on content learned several years earlier by most individuals, content presumably equally familiar to almost everyone. Achievement test items, on the other hand, are more circumscribed. They are necessarily drawn from the restricted subject-matter of a particular course of training--in chemistry, European history, and Latin, for example--usually a recent course.

5. Inasmuch as achievement tests are based on a relatively narrow domain, known and understood best by those who have been exposed to that domain, they (obviously) cannot be used for evaluating the educational outcomes for individuals who have not been exposed to it. Aptitude tests, however, draw from much wider domains, not confined to the material learned in classroom, and are presumably within the actual, or accessible, experiences of all individuals. Therefore, unlike achievement tests,

aptitude tests can be used to make general intellectual evaluations for all who share a common culture regardless of their particular classroom experiences. On the other hand, because their coverage is not classroom-specific, aptitude tests cannot be used, as achievement tests can, to evaluate the quality of particular educational programs.

6. Aptitude is by its nature prospective--indeed the word "aptitude" itself has implications for the success of future learning--and scores on an aptitude test are typically used for predicting future success in the general domain of that aptitude. Not only is the sense of aptitude prospective, it sometimes implies that the learner whose aptitude is being evaluated has not yet been exposed to the subject-matter to be learned and therefore cannot yet be tested on it. Achievement is by its nature retrospective--also implied by the word--and achievement tests are typically used to evaluate the level of accomplishment in prior learning experiences. This is not to say that achievement tests cannot or have not been used to predict future success. They have, and they are very useful for that purpose. Past achievement is always a good predictor of future achievement, indeed often a better predictor than aptitude scores.

In spite of the foregoing, the distinctions between aptitude and achievement are not entirely clear. Aptitudes are necessarily, in some sense at least, developed abilities (Green, 1978), albeit much more rapidly and thoroughly developed in some individuals than in others. Therefore, it should be understood that despite the foregoing distinctions, aptitude tests are, fundamentally, also achievement tests (which, clearly, also measure developed abilities), but tests that are not dependent on a specific curriculum. But even this distinction is not absolute. It is true that many aptitude tests, like the SAT and the GRE General Test, make use of some school-learned verbal and mathematical skills. What helps to justify the claim that these are not achievement tests in the usual sense is that the concepts tested are meant to call for generalizations, inferences, and special insights that go beyond the specifics of the subject-matter originally studied.

Further, some of the distinctions between the two constructs are virtually impossible to validate empirically--for example, that aptitude develops outside the school environment as well as inside it. Other distinctions are researchable, such as the resistance of aptitude to educational interventions after learning patterns have been established in childhood; and, in fact, much investigative work has been carried out in this connection.

As has already been suggested, a frequent difficulty in working with the aptitude-achievement distinction is the tendency to confuse the construct with the measure of the construct. In most instances, it is easy to identify a test as an achievement test; tests consisting entirely of chemistry items, history items, philosophy items, physics, or French, for example, are clearly achievement. Items of reading comprehension or vocabulary or quantitative problem solving, however, which are often used in aptitude tests, are sometimes also used in constructing achievement tests, a practice that, while understandable, does tend to contribute to the confusion. For various reasons our tests of aptitude and our tests of achievement are often seen to be measuring quite similar abilities. We find, for example, that the correlations between SAT-verbal scores with the College Board Achievement Test scores in English Composition and in English Literature are in the low to middle .80s. The correlations between SAT-mathematical scores and scores on the Achievement Tests in mathematics (Mathematics Level I and Mathematics Level II) are similarly in the low .80s (Donlon, 1984). Although the correlations between the SAT and other Achievement Tests in the battery are lower than .80, some considerably lower, the correlations just cited are probably higher than we would feel is ideal for pairs of tests that are thought to be measuring different constructs.

It should be noted, in passing, that the foregoing correlations apply to situations in which the aptitudes are almost fully developed, but where achievement is not. It is possible that these relationships might take on different patterns when both are undergoing change, as in childhood. On the other hand, this latter effect may not be easily ascertainable; the distinctions between aptitude and achievement are more difficult to demonstrate and measure at early stages of development.

Nevertheless, in spite of these inadequacies in the measures we have constructed, many (e.g., Bereiter, 1974, and Carroll, 1974) would argue that the construct of academic aptitude "deserves a conceptual status distinct from achievement" (Bereiter, 1974), and should not be abandoned simply because of the confusions and tensions we have experienced in defining it. The same confusions, one might argue, are present in the definitions of other types of aptitudes. It does suggest, however, that we must continue to search for measures that are distinctly different from achievement, items that focus more on process than on content, and items that vary in difficulty and discriminate over a wide range of talent but depend on material learned only at elementary levels.

The concept of aptitude seems also to have suffered from its association with the nature-nurture controversy and with the assumption that characteristrics that are inherited and "innate" are firmly resistant to change at <u>any</u> age. Why this view is held is hard to say. We know of several genetically determined physical disorders--phenylketonuria, galactosemia, hemophilia, and diabetes, for example--that are quite responsive to environmental interventions, and many others--stature, for example--that have long been known to be changeable over generations, probably as a function of changing diet.

The converse of this view seems also to be held: that inasmuch as aptitudes are frequently in continuous change, they cannot be innate. That they are in continuous change especially during the very early years, cannot be denied; raw scores and some types of scaled scores on aptitude and intelligence tests grow rapidly during that time. Even the claim of IQ constancy is an implicit admission that mental ability changes, but that the change is indexed to the change in chronological age. But one does not follow from the other; change in the level of aptitude is not by itself evidence that it is not innate--and there is considerable evidence that aptitude, or intelligence, has a large genetic component. As already indicated, many characteristics that are known to change are also acknowledged to be innate (and vice versa), even within a lifetime: stature (again), arm length, and hirsuteness, for example, and most other physical characteristics. The genetic pattern is laid down at the time of conception, but the characteristics themselves change continuously, sometimes not even appearing until later in life, often not until adulthood.

Thus, it appears that the issue for useful consideration is not whether aptitude is innate; indeed the issue of innateness is irrelevant in the present context. Nor is the fact of ordinary change, i.e., predictable change associated with change in age, a useful issue in this context. What is at issue is whether there can be <u>differential</u> change in the individual, that is, whether and to what extent differential environmental experiences, including special intervention strategies, can exert a differential impact on scores. Currently, the view is that within the normal range of intelligence, aptitudes are indeed susceptible to differential cognitive training, but that the training must begin very early in life and continue for an extended period through the formative years and beyond; further, that the cognitive training must be carried out in a continuously supportive and motivating atmosphere.

It has already been pointed out that what makes the concept of aptitude particularly difficult to deal with objectively is its implications, as some see it, for the present and future status of minority groups in our society. The thesis here is that there is no justification for such implications. But in order to understand better the mechanisms that are characteristic of aptitude and what they do imply, other urgent questions have developed--whether, for example, scores on aptitude tests rise or fall differentially as a function of ordinary intervening experience, in particular during the period of early adulthood.

It is to this latter question that the present study is addressed: Given a sample of students classified by sex and undergraduate field of study--humanities, social science, biological science, or physical science--to what extent does the rank order of these students on verbal and mathematical aptitude tests change over the period of time in which they are enrolled in college? Second, what are the differences in aptitude test scores, (verbal, quantitative, and analytical) among students of different sex and field of study, after controlling on initial score? This question, which is most particularly addressed to the matter of differential impact of curriculum on aptitude scores, may be stated as follows: Given two students of equal ability, as evidenced by their SAT scores, one who majors in the humanities area in college, the other, in the physical sciences. Will the first, after four years, earn higher scores on the GRE-verbal Test than the second, and will the second earn higher scores on the GRE-quantitative Test than the first, and by how much? What will be the impact on the GRE-analytical scores? And to what extent is the sex of the student a determining factor in these differences?

There are several questions to be investigated in the course of these analyses. One, already alluded to, is the extent to which verbal, quantitative, and analytical aptitude test scores on the GRE General Test are affected by the student's gender and/or educational exposure to one or another major field of study. A related question is: To what extent are differences in initial aptitude test scores critical in producing differences in later aptitude test scores; and how do these differences vary as a function of sex and field of study? Second, are these effects heightened if we confine our study to the more clearly "verbal" and more clearly "quantitative" fields of study? Third, on the presumption that a particular curriculum studied may vary sharply in content and in level of demand from one college to another, would the results of the study be altered in any significant way if the outcome scores were conditioned also on college attended? Finally, the question is asked, is it possible to identify aptitude items that are more affected than others by sex and intervening academic experience, and to characterize them in a way that will provide guidance in the development process?

FORMATION OF THE STUDY SAMPLE

The population of interest for the study was conceived of as consisting of those who took the SAT and also the GRE General Test at the normal times in their academic careers, with the typical number of years intervening. Accordingly, the database for the study was defined by first selecting all examinees who took the GRE General Test, Form 3FGR2, in October 1983, April 1984, October 1984, and February 1985, and all who took Form K-3FGR3 in December 1984. From this total group only college seniors were selected, yielding a total of about 34,000 cases. The list of these students was then compared with the file of SAT takers four and five years earlier and a matched sample of students taking both tests was assembled, including students who had taken the SAT as high school juniors or seniors <u>and</u> the GRE as college seniors. These cases were further examined to confirm that information on sex and undergraduate major field of study was available and was further reduced to include only those for whom English was their primary language at the time they took the SAT. When the study sample was finally assembled, it consisted of a total of 22,923 cases, of whom 12,601 had taken Form 3FGR2 and 10,322 had taken Form K-3FGR3 of the GRE General Test.

The total sample was subdivided for study purposes by sex and undergraduate major field of study, as defined by the 1984-85 GRE Bulletin (Educational Testing Service, 1984; see the Appendix), yielding the following numbers in each cell:

	Men	Women	Total
Humanities	1,305	2,141	3,446
Social Science	3,031	5,514	8,545
Biological Science	1,561	2,969	4,530
Physical Science	4,626	1,776	6,402
Total	10,523	12,400	22,923

Finer breakdowns than those given above may be useful in considering the results of the analyses. Table 1 gives counts of the study sample by ethnic background, field of study, and sex. Close examination of Table 1 will reveal that the counts by major field differ quite considerably across the ethnic groups. For example, Blacks are heavily concentrated in social science, but underrepresented in the other three fields. Hispanics are somewhat overrepresented in social science but very much underrepresented in physical science. The numbers of Hispanics enrolled in the humanities and biological science, however, are about what would be expected on the basis of the total numbers in those particular fields across all ethnic groups and on the basis of the total number of Hispanics across all major fields. Asians, as expected, are heavily concentrated in physical science, but relatively sparse in social science and only slightly so in humanities.

As expected, the men are overrepresented in physical science and underrepresented in the other three fields. Conversely, and also as expected, the women are underrepresented in physical science and overrepresented elsewhere. Leaving aside the physical science area, however, the two sexes are distributed in about the same proportions in the remaining three areas.

RESULTS

Review of Summary Statistics, by Subgroup

The data in Tables 2a to 2i describe the intercorrelations, means, and standard deviations among the five variables of interest--SAT-verbal, SAT-mathematical, GRE-verbal, GRE-quantitative, and GRE-analytical--for the total sample of 22,923 cases and for the eight component subgroups of the total, broken down by field of study and sex. Table 2j provides a convenient summary of the numbers of cases, the means, and the standard deviations for all the subgroups of the study sample. Focusing on the total sample for the moment, we observe that it is a highly select subgroup of the typical SAT population, yielding a mean of 519 on SAT-verbal, 94 points higher than the corresponding mean of 425 for the entire candidate population tested in 1986-87 (Educational Testing Service; October, 1987), the most recent year for which such data are available. The study sample also shows a mean of 556 on SAT-mathematical, 84 points higher than the SAT-mathematical mean of 472 for candidates tested in 1985-86. It is also noted that its standard deviations of 105 on Verbal and 110 on Mathematical are slightly lower than the standard deviations of the 1986-87 reference population--106 and 118, respectively--suggesting the fact of their selectivity. The sample also appears to be selective in terms of the GRE population. Its means of 510 on GRE-verbal, 573 on GRE-quantitative, and 580 on GRE-analytical are higher than the means of 493, 553, and 546, respectively, for seniors and nonenrolled college graduates who took the GRE between 1983 and 1986 (Educational Testing Service, 1987-88; 1987, p. 15). Its standard deviations, of 108 on GRE-verbal, 126 on GRE-quantitative, and 118 on GRE-analytical are lower than those of the reference population just cited, namely 118, 132, and 125, again pointing to the selectivity of the study sample, even in relation to the GRE population of seniors and nonenrolled college graduates who are themselves a select subgroup of the total GRE candidate population.

The foregoing findings are not overly surprising, however, in view of the fact that the members of the study sample were not expected to be typical of the general SAT population. These students, unlike the SAT population whose plans may or may not call for further education beyond the bachelor's degree, are all applying for admission to graduate school, and should therefore be expected to be a higher-scoring subset of the SAT population.

What is particularly interesting about the data in Table 2a (which are based on the entire study sample of 22,923) in the context of the present study are the correlations between SAT-verbal and GRE-verbal and between SAT-mathematical and GRE-quantitative, both of which are .86, indicating that there is a substantial linear relationship between SAT and GRE scores that explains virtually three-quarters of the variance in GRE-verbal and GRE-quantitative scores taken four years later. It is recalled that these students are quite diverse with respect to their academic interests, having gone their separate ways after high school into a wide variety of college majors, where their verbal and mathematical skills would be expected to undergo differential change. It is therefore particularly interesting that, over all, their rank order in these two general aptitude areas at the time of their junior or senior year in high school has been so well preserved.

The pattern of correlations with GRE-analytical are also of some interest. We note that the correlations of GRE-verbal and GRE-quantitative with GRE-analytical are of the same magnitude, respectively, as the correlations of SAT-verbal and SAT-mathematical with GRE-analytical. In such comparisons we note that the mathematical and quantitative correlations with analytical are higher than the correlations of verbal with analytical. We note further that each of these several correlations is higher than the ver-mathematical or verbal-quantitative correlations, but lower than the .86 correlations of verbal with verbal and mathematical with quantitative that were noted above. These data would suggest that the GRE-analytical test is a composite of verbal and mathematical material and are supported by other data (e.g., Educational Testing Service; April 1985, June 1985), in which we learn that indeed these patterns of correlations with GRE-analytical come about because of the composite structure of that test. The test consists of two item types, Analytical Reasonsing and Logical Reasoning, in a ratio of number of items of about 3 to 1. The former of these two groups of items correlates more highly with GRE-quantitative; the latter, more highly with GRE-verbal.

Ordinarily, it is customary to discuss differences in means before going on to discuss measures of variability. In this case, however, the usual order will be reversed; a detailed study of the means of these groups can be gleaned best from tables (3a to 3e) that appear somewhat later in this report.

As expected, the individual subgroups are generally more homogeneous than the total group (See Tables 2a to 2i, and 2j), although there are some exceptions, mostly in the case of the verbal tests. In the case of both the SAT-mathematical and the GRE-quantitative Tests all subgroup standard deviations are smaller than the total-group standard deviations, some by substantial amounts. On the GRE-analytical Test six of the eight subgroups (exceptions: men in the humanities and in social science) show standard deviations smaller than that for the overall total group.

The data that follow in this section of the report will attempt to describe the nature and degree of the differential impact of their sex and college curriculum on their GRE aptitude test scores. Before going on to the analysis of impact, however, it may be useful to compare the means on the five variables of interest across the eight subgroups. Tables 3a to 3e correspond respectively to the SAT-verbal, SAT-mathematical, GRE-verbal, GRE-quantitative, and GRE-analytical Tests, (and summarized in Table 2j). Each table presents, for the specified test, the mean scores by sex within field of study. Also presented within each field of study, are the (unweighted) average of the mean scores for men and women and the differences in mean scores between men and women. The values of the averages of the male and female means address the question of whether there is an average difference in performance by field of study, irrespective of sex. The values of the male-female differences in means within field of study address the question of whether there is a difference in performance between the two sexes and whether this difference, if it exists, is associated with a particular field of study.

It should be noted that the averages just referred to are unweighted averages and are for that reason better suited for the purpose of this comparison than the simple averages within field of study across all students, since the unweighted averages remove the confounding effects of the differential representation of the sexes within field of study. (For example, the simple average of scores of all students in physical science is $.72\ X_{MP} + .28\ X_{FP}$, where X_{MP} and X_{FP} are the mean scores for men and women, respectively, in which the coefficients of the means represent the relative numbers of men and women. Similarly, the simple average of scores of all students in humanities is $.38X_{MH} + .62X_{FH}$. The resulting difference between these two simple averages largely compares the performance of men in physical science with women in humanities. Consequently, it includes (inappropriately here) a component of any consistent difference in performance, across field of study, between the sexes. The unweighted averages do not suffer from this confounding).

Tables 3a to 3e also include standard errors of each of the statistics presented as well as two measures of the potential difference in performance between the 8 subgroups. One measure is the F-statistic from a standard one-way analysis of variance. Because of the large sample sizes, this statistic, which has 7 and 22,915 degrees of freedom, is best suited for the comparison of the way in which the ratio of the between-groups variance to the within-groups variance changes across the various aptitude tests under consideration. A better measure of the extent that a student's performance depends on subgroup membership is:

$$ETA^2 = 1 - SS_W/SS_T,$$

where SS_W is the pooled within-subgroups sum-of-squares and SS_T is the across-subgroup (total) sum-of-squares. ETA^2, which is analogous to R^2 in regression, measures the proportion of the total variability of the test score that can be accounted for by taking subgroup membership into account.

Upon examining Tables 3a and 3c we see that, as expected, the verbal means on both SAT and GRE are highest for the humanities groups. Of special note is that, except for the SAT-verbal scores for the humanities group, the scores of men on both verbal tests are higher than those of women within the same field of study. In years past this was not so; the mean scores for women exceeded those for men by about 6-7 points. In recent years, however, this difference appears to have been reversed; men's scores exceed the women's now, by at least that amount. What is also of some interest is that the physical science groups are not far behind the humanities groups on the verbal tests. There appears to be some interaction between sex and field of study, but only on the SAT. There the women outscore the men in the humanities area; in all the other fields the men outscore the women. On the GRE-verbal the men outscore the women in all the fields. The lowest-scoring of all eight subgroups on both the SAT-verbal and GRE-verbal is the female social science group, followed closely by the female biological science group.

On the quantitative side (Tables 3b and 3d), the highest-scoring by far are the physical science groups, with the men scoring substantially higher than the women, confirming the observation made in virtually every other such compilation of quantitative data, in which it is found that physical science groups outscore all other groups by a considerable margin, and where the men consistently outscore the women. At the other end of the scale we find here that, as in the verbal tests, the social science field is the lowest-scoring on the quantitative tests, with (again) the men outscoring the women. In fact, the difference in performance between men and women is relatively constant across fields of study, and, as just indicated, quite consistent with virtually all such tabulations reported in the literature.

Scores on the GRE-analytical Test (Table 3e) follow the quantitative pattern for the most part, with the physical scientists in the clear lead, followed at some distance by the humanities group. Again, the women in each curriculum group follow the men on the analytical test, but not at quite the same distance as on the quantitative test. The differences in mean scores by sex within field of study resemble those on the verbal tests with the difference being largest for the social and biological sciences. As in the verbal and quantitative tests, the social science group is the lowest scoring on the GRE-analytical of all the major fields.

It may also be useful to see graphically the disposition of the eight subgroups in terms of their bivariate means in both the verbal and quantitative domains and the manner in which the groups display themselves along the outcome (GRE) measure. Figure 1 is a schematic plot of the nine bivariate means of SAT-verbal vs GRE-verbal, one for the total study group and one for each of the eight component sex x field-of-study groups. Figure 2 is the same sort of picture for SAT-mathematical vs GRE-quantitative.

The differences between Figure 1 and Figure 2 are noteworthy. In Figure 1 the eight subgroups and their bivariate means are closely clustered, showing very little dispersion along the main diagonal and little differential effect of either the factor of sex or the factor of field of study. Figure 2, on the other hand, evinces much more dispersion than does Figure 1 along the main diagonal shown by the bivariate means; and contrary to appearances, it also shows 2.5 times as much dispersion in the off-diagonal direction than is true of Figure 1. For example, the bivariate means (centroids) for the male and female humanities groups are displaced downward from the general line of the means while the centroids for the male and female physical science groups are displaced upward (as well as being very high on both SAT-mathematical and GRE-quantitative), suggesting that the latter groups are higher-scoring on GRE, relative to SAT, than are the groups of men and women in the humanities. The former are relatively lower-scoring, even in relation to their earlier SAT-mathematical scores. This, in turn, suggests that the college mathematics curriculum had a positive impact on the GRE-quantitative scores of the physical science majors. The GRE-quantitative scores of the humanities majors, who in all probability generally took little or no mathematics in their college years, suffered in comparison to the other groups. More detailed analyses of this phenomenon appear later in this section of the report.

It will also be useful to discuss an apparent contradiction in the results just described, that is, that the correlation between SAT-mathematical and GRE-quantitative for the entire group of 22,923 is no lower--indeed, very slightly higher--than the correlation between SAT-verbal and GRE-verbal (also given for the entire group), despite the fact that there is so much greater differential impact in the mathematical-quantitative domain (see Figures 1 and 2). The reason for this is that the groups, as indicated above, are not only more diverse in the off-diagonal direction on the mathematical than on the verbal tests, they are also much more diverse along the main diagonal defined by the centroids. The range of SAT-verbal means is 58 points, extending from 493 to 551 and the range of GRE-verbal means is 67 points, extending from 481 to 548. In sharp contrast, the ranges of mathematical means are more than twice the ranges of verbal means. The range of SAT-mathematical means is 140 points, from 500 to 640; the range of GRE-quantitative means is 187 points, from 499 to 686. A more precise description of this phenomenon can be made in terms of ETA^2, the ratio of between-groups variance to total (over group) variance on each of the four measures. The values of ETA^2 (from Tables 3a to 3e) are .052 for SAT-verbal and .050 for GRE-verbal, which are markedly smaller than .217 for SAT-mathematical and .308 for GRE-quantitative and indicate that the standard deviations of mathematical-quantitative scores in the individual subgroups are uniformly smaller than they are in the total group. This is much less the case for the verbal scores.

On the other hand, it is at least barely possible that the low curricular impact on GRE-verbal is a function of the nature of the items that constitute that test. Each form of the GRE-verbal Test is balanced so as to include about equal numbers of items from the humanities, the social sciences, and the physical and biological sciences. Conceivably, the differential impact of curriculum might be more clearly visible if the items of the test were confined to one or another of these domains, rather than a balance of all four.

Analysis of the linear relationships between SAT and GRE scores

We have noted above that there are substantial relationships between scores on the SAT and scores on the GRE. In particular, we have observed that nearly 75% of the total variability of the GRE-verbal and the GRE-quantitative scores can be accounted for by simple linear regressions of those scores on, respectively, the scores on SAT-verbal and SAT-mathematical. Consequently, much of the information about how a student's field of study might differentially impact that student's GRE aptitude test scores, after controlling for SAT scores, can be obtained by examining how the linear relationships between GRE and SAT scores vary by subgroup of student.

In the initial phases of the study regression analyses were carried out between GRE scores and items of information called for on the questionnaire that students are asked to fill out at the time they take the SAT. Such items include mother's and father's educational level, the student's rank in class, type and amount of study and grades in various subjects, educational plans, etc. Responses to these items, along with the SAT scores, were included in multiple regression equations in an attempt to improve the prediction of the GRE scores. However, there was great variation among the eight groups with respect to the kinds of variables that would improve prediction beyond what was already possible with SAT-verbal and -mathematical, and in no case was the multiple correlation raised by any significant amount. Therefore, in an effort to standardize the prediction variables across the eight groups, it was decided that throughout the study we would use only SAT-verbal and SAT-mathematical as predictors. It is noted, however, that even these variables failed to behave uniformly. Although the addition of SAT-verbal to SAT-mathematical did aid in the prediction of GRE-analytical, the addition of SAT-verbal to SAT-mathematical helped only negligibly in the prediction of GRE-quantitative.

In any case, the study of differential impact depended on the use of only the SAT-verbal and SAT-mathematical as control variables. Consequently, the examination of the way in which the linear predictive relationship varies by subgroup of student will be the thrust of the present section.

We will begin with the linear relationships, by subgroups, predicting scores on the GRE-verbal from scores on the SAT-verbal alone and predicting scores on the GRE-quantitative from scores on the SAT-mathematical alone. Table 4 shows the result of fitting the model,

$$\text{GRE-V} = a + (\text{SAT-V})b,$$

separately within each of the eight sex-by-field-of-study subgroups of students. In addition to providing the values of the intercept and slope of the within-group regressions, Table 4 also includes the standard error of estimate, the value of R^2, and the amount that the value of R^2 could be increased by adding the student's SAT-mathematical score to the prediction equation. We see (in the column headed R^2) that between 69 and 75 percent of the total variation of the GRE-verbal scores within any group can be accounted for by the within-group simple linear regression on SAT-verbal score and (in the last column) that the inclusion of the SAT-mathematical score adds little additional information, increasing the explained variation by at most one percent. We see that the equations for men and women within the same field of study tend to resemble each other although the slopes for the women are slightly flatter than those for the men, suggesting that differences in SAT-verbal scores are less critical in predicting GRE-verbal scores for women than for men. Furthermore, the slopes for students of either sex in humanities are noticeably steeper than those in any of the other fields and the slopes of students in the biological science are noticeably flatter than those in any of the other fields. This suggests that, in predicting GRE-verbal scores, SAT-verbal scores are most critical for the humanities students and least critical for the students in the biological sciences.

Table 5 shows the results of the within-group regressions of GRE-quantitative score on SAT-mathematical score. The linear relationship between the two scores is fairly strong, almost as strong as in the case of SAT-verbal and GRE-verbal, accounting for between 63 to 73 percent of the total variation in GRE-quantitative scores. Second, we see here, in contrast to Table 4, that, with the exception of the social science majors, the slopes for women are generally steeper than those for men, suggesting that differences in SAT-mathematical scores are more critical in predicting GRE-quantitative scores for women than for men. We see also that the range of the slopes for the various subgroups of students is much larger here than it was for the verbal aptitude test scores. Finally, it is apparent (as indicated above) that little information in terms of predictive power can be gained by including the SAT-verbal score in the model, even less than by including SAT-mathematical in predicting GRE-verbal scores.

Our goal is to examine how the relationship between GRE and SAT scores varies over subgroups. In order to do this, it is more informative to compare the entire regression lines rather than the within-group slopes alone, as shown in Tables 4 and 5.

Figure 3 shows the eight within-group regression lines for the prediction of GRE-verbal score from SAT-verbal score and Figure 4 shows the corresponding lines for the within-group prediction of GRE-quantitative score from SAT-mathematical score. The most striking observation in comparing the two figures is that the prediction lines for verbal aptitude are much closer together than are the prediction lines for quantitative aptitude. Of additional note is the fanning of the quantitative aptitude prediction lines for lower levels of SAT-mathematical aptitude. The interpretation of these observations is that the between-group variability of predicted GRE scores for given ability is greater for the measure of quantitative aptitude than for the measure of verbal aptitude, and particularly so for the lower levels of ability. In other words, the differential impact of field of study on aptitude test scores is greater for the quantitative than for the verbal aptitude measures, especially so at lower levels of ability. This is an observation that is made several times in reviewing the data summarized in this report.

The lines shown in Figures 3 and 4 convey the main information about the characteristics of the linear relationships between GRE and SAT aptitude test scores. However, because of the clustering of the constituent lines in Figure 3, there is not enough resolution to allow us to assess conveniently how these linear relationships vary by subgroup.

Predictions of GRE scores from SAT scores, by subgroup

We would like to determine if the relative standing, as measured by predicted GRE score, of each of the eight subgroups is different for different levels of SAT and, if so, by how much. One way to address this question is to examine how the predicted values of GRE-verbal score (for example) for a specified value of SAT-verbal score depend on the sex-by-field-of-study subgroups. Table 6 does this for the predicted values of the GRE-verbal score for two extreme levels of initial ability, corresponding to SAT-verbal scores of 380 and 650, the scores at the 10th and 90th percentiles of the distribution of SAT-verbal scores for the total study sample. Besides the eight subgroup mean predicted values for each of the two levels of verbal ability, the table includes the average of the scores for men and women within each field of study, the within-field-of-study differences between the scores of the two sexes, and the standard errors of all statistics. The last column of the table gives the ranges of predicted scores, averages, and differences across the four major fields of study.

In a similar manner, we can compare the predicted values of the GRE-quantitative score for two equivalently extreme levels of initial mathematical ability, namely, 400 and 700, corresponding, respectively, to SAT-mathematical scores at the 10th and 90th percentiles of the distribution of SAT-mathematical scores from the total study sample. The result is shown in Table 7. (Note that, since SAT-mathematical scores tend to be higher than SAT-verbal scores, the 10th and 90th percentiles for the SAT-mathematical scores are, at 400 and 700, somewhat higher than those of the equivalent percentiles for the SAT-verbal.)

It is interesting to observe in Tables 6 and 7 that even when the data are conditioned on SAT scores, the GRE means for the men are higher than those of the women in both verbal and quantitative, and that this observation is consistent across all four fields of study, at both the low and high score levels on SAT. In the case of GRE-quantitative the difference in predicted means is probably to be expected; men generally take more quantitatively-oriented courses in college than women, and this greater exposure to mathematics is likely to raise their GRE-quantitative scores beyond those of the women, even for men and women who have earned the same initial (SAT) score. As expected, the highest predicted GRE-quantitative scores are for men in the physical sciences (see Table 7)--510 for an initial SAT-mathematical score at the 10th percentile and 730 for an initial SAT-mathematical score at the 90th percentile. What is harder to understand are the higher predicted means in GRE-verbal for the men than for the women (see Table 6), with differences in favor of the men averaging about 12 points for students with SAT-verbal scores at the 10th percentile and about 18 points for students with SAT-verbal scores at the 90th percentile. These differences are not substantially different from the corresponding predicted mean differences between men and women on quantitative, in which there are differences of about 28 for SAT-mathematical scores at the 10th percentile and about 16 for SAT-mathematical scores at the 90th percentile.

As an adjunct to these tables of predicted values for extreme levels of initial (SAT) aptitude, we can provide a graphical display of how the relative predicted performance of the subgroups change as the value of initial aptitude changes by adjusting the plots in Figures 3 and 4 to remove the overall estimate of the linear relationship between GRE and SAT score. Accordingly, we have done this in Figure 5 for the case of the verbal aptitude test scores. Each line in this plot corresponds to one of the eight subgroups and is the difference between the prediction line for that subgroup from Figure 3 and an average line describing the across-group relationship between GRE-verbal and SAT-verbal test scores. (This average line is defined by $y = \bar{a} + x\bar{b}$ where $\bar{a}$ and $\bar{b}$ are, respectively, the unweighted averages of the within-group intercepts and slopes). Thus these lines show, for any given SAT-verbal score, and for each subgroup, the difference between the predicted value of GRE-verbal score for that subgroup and the overall average predicted score across all subgroups. Similarly, Figure 6 shows, for each subgroup and for each value of SAT-mathematical score, the predicted values of GRE-quantitative score relative to the overall average predicted score across all subgroups.

Upon examining these tables and figures we see again that the most striking difference between the predictions of verbal and quantitative measurements of aptitude lies in the between-group variability in predicted GRE scores for students with lower SAT ability. The range of subgroup mean predicted GRE-verbal scores averaged across both sexes, assuming an initial SAT-verbal score at the 10th percentile score of 380, is 14, with ranges of predicted scores within sex of 15 for men and 12 for women. The corresponding range of subgroup mean predicted GRE-quantitative scores, also assuming an initial SAT-mathematical score at the 10th percentile score of 400, is about six times as large--91 points (with correspondingly larger within-sex ranges of 97 and 84, respectively, for men and women.) The ranges of predicted subgroup mean scores at the other end of the scale of initial abilities (SAT scores at the 90th percentile--650 on verbal and 700 on mathematical), also averaged across both sexes, are much closer together. These predicted scores are 17 points for GRE-verbal and 43 points for GRE-quantitative. (The within-sex ranges are 17 and 18, respectively, for men and women on verbal, and 39 and 47, respectively, for men and women on quantitative.)

The ordering of the subgroups in terms of their predicted GRE-quantitative score is consistent for both values of SAT-mathematical score and is quite suggestive. For any given level of initial performance, men and, to a lesser extent, women in the physical sciences are predicted to perform at a noticeably higher level than students at the same level of initial ability in any other field of study. The ordering of the remaining subgroups of students, in terms of their predicted quantitative score, given any initial mathematical score, is in the same direction as the probable exposure to heavily quantitative coursework. We will present some interpretations for this ordering shortly. It appears fairly clear that there is a differential impact of field of study on the quantitative score and that the quantitative findings are generally consistent with expectations.

Less clear is the relationship between subgroup membership and predicted verbal score (see Table 6). For example, of all the students with SAT-verbal scores at the 10th percentile, the lowest predicted GRE-verbal means are those for the humanities groups, and especially so for women in humanities. (These predicted GRE-verbal means are essentially the same as those for the physical science groups). The ordering of the subgroups in terms of their predicted GRE-verbal score depends on the initial SAT-verbal score (as may be seen from Figure 5), with an exchange in relative position occurring at around 500. Above this value the subgroups are roughly ordered, within sex, in approximate relation to the content of verbal material in their coursework, so that humanities majors, as would be generally expected, are predicted to score higher than students of the same sex who have majored in other fields. Below 500, differences among fields show no clear or rational pattern, but the predicted scores appear to be somewhat lower for humanities and physical science than for social or biological science. Further, it is important to emphasize that (as pointed out above) the predicted scores are in every field and at every level higher for men than for women.

Because we have been basing our comparisons of subgroups on predicted values determined by lines fit through the data, it will be useful to see how far those linear predictions diverge from the actual values. Table 8 shows the means and standard deviations of GRE-verbal scores, by subgroup and by each of the following four ranges of SAT-verbal score: 351 to 450, 451 to 550, 551 to 650, and 651 to 750. Correspondingly, Table 9 shows the means and standard deviations of GRE-quantitative scores, also by subgroup and by the same four ranges of SAT-mathematical score. Corresponding to these tables are Figures 7 and 8 which show, by subgroup and range of SAT-score, the mean residuals (actual minus predicted) from the linear predictions of GRE-verbal and GRE-quantitative scores, respectively. The main impression from Figure 7 is that there is a consistent and roughly quadratic nature to the plots of the mean residuals from the prediction of GRE-verbal score, indicating that the scores for students with low and high initial SAT-verbal scores will be underpredicted while the scores for the moderate performers will be overpredicted. The other point of note from the figure is that this effect is small and fairly consistent across subgroups (with the possible exception of males in the social sciences). The residuals from the prediction of GRE-quantitative score from the SAT-mathematical score shown in Figure 8 are of small magnitude and display no observable pattern.

We turn now to the GRE-analytical test and examine how the linear relationships between scores on that test and scores on the SAT tests depend on field of study and sex. Table 10 presents the coefficients from the within-group regressions of GRE-analytical score on both SAT-verbal and SAT-mathematical scores. We see from the table that the linear prediction of GRE-analytical score from the SAT scores in less strong than the predictions of GRE-verbal and GRE-quantitative, accounting for between 52% and 58% of the total variability of GRE-analytical scores, as compared with ranges of 69% to 75% for verbal and 63% to 73% for quantitative (see Tables 4 and 5). We also see that both SAT-verbal and SAT-mathematical scores are required in the equation although the SAT-mathematical score is consistently the more important predictor.

As was the case for the GRE-verbal and the GRE-quantitative tests, a sense for the way in which the relationship between GRE-analytical scores and SAT scores varies over subgroups can be obtained by comparing the within-group predictions of the GRE-analytical score for given values of the SAT-verbal and the SAT-mathematical scores. Because these predicted scores depend on both the SAT-verbal and the SAT-mathematical scores so that each of the within-group predictions describes a plane, the direct graphical representation in two dimensions of the predicted scores for all values of the SAT-verbal and the SAT-mathematical scores is problematical. For graphical convenience, and to produce prediction lines roughly comparable to the prediction lines used for the GRE-verbal and GRE-quantitative tests

(Figures 3 through 6), we will consider the within-group prediction lines of the GRE-analytical score for each measure of SAT ability as defined by equal scores on both the verbal and mathematical tests. The result is shown in Figure 9. To obtain better detail, we subtract out the average of the eight within-group prediction lines to produce Figure 10. We see that the preuiction lines are clustered and appear generally parallel.

Before further considering the relationships between subgroup membership and predicted GRE-analytical score, it is necessary to observe that the lines plotted in Figures 9 and 10 pertain to students who are relatively more proficient on the verbal scale than they are on the mathematical scale. This is because the lines assume equal scores on both tests, even though the SAT-mathematical scores tend to be higher in our sample than the SAT-verbal scores so that, for example, a score of 650 corresponds to the 80th percentile on the SAT-mathematical test while the same score of 650 is near the 90th percentile of the SAT-verbal test. To place the initial verbal and mathematical abilities on more equal footing in the prediction of the GRE-analytical score and to allow comparison with the GRE-verbal and GRE-quantitative results, we present Table 11, which shows the predicted scores for low and high initial ability students, defined respectively as having both verbal and mathematical SAT scores equal to the 10th and 90th percentiles of their respective score distributions. (Thus, low ability students have initial SAT-verbal and SAT-mathematical scores of 380 and 400, respectively, and high ability students have scores of 650 and 700, respectively. It should be noted, in passing, that the lines formed by connecting the predictions in Table 11 for the low initial ability students with the predictions for the high initial ability students are very nearly the same as the lines shown in Figures 9 and 10).

It is interesting to observe, in Table 11, that the range of predicted scores across the fields of study, both within sex and averaged across sex for each of the two initial ability levels is relatively small, closely resembling the differences for the observed GRE-verbal scores (Table 6). Furthermore, the ranges of GRE-analytical scores for SAT scores at the low initial ability level are much smaller than the corresponding ranges of the GRE-quantitative scores at that SAT score level (Table 7). It is also interesting to observe that, unlike the predictions of verbal and quantitative scores, the predicted GRE-analytical scores of women are consistently higher than those of their male colleagues in the same field of study. This is so even though their unconditioned means on the analytical test are lower than those of the men.

In summary (and as indicated above), the differential impact of field of study on aptitude as measured by the GRE tests appears to be much less for both the verbal and analytical tests than for the quantitative test and the differential impact for the quantitative test, especially across fields of study (not as much across sex) is greater for students of lower initial ability than for students of higher initial ability.

Since the GRE-quantitative scores of the physical science majors appear to be most heavily impacted by their previous mathematical training, it may be helpful to use them for illustrative purposes in examining their self-selective characteristics.

Observe that the sample of students that we have been studying, while very large, is self-selected, consisting only of those students who are planning to pursue graduate education--in most instances, within the same general field of study as their undergraduate education. Successful graduate study in the physical sciences requires a certain minimum level of mathematical ability. It would appear that the students in our sample who were majoring in the physical sciences and who had lower initial mathematical ability (as measured by their SAT-mathematical score) nonetheless presumably believe, at the close of their undergraduate career, that they are capable of pursuing graduate study and that they have achieved the necessary level of mathematical competence. Not included in our sample are the colleagues of these students, those physical science majors of lower initial ability who did not achieve the level of mathematical ability necessary to pursue graduate study and so declined to take the GRE. Consequently, our sample of physical science majors may consist largely of those students who feel that they have achieved a minimum level of mathematical ability, some of them, perhaps, in disregard of their low initial measure of mathematical ability at the time they took the SAT.

Coupled with this self-selection is the possibility that the tasks required by the quantitative test are more likely to be impacted by experiences in the classroom, specifically in courses in the physical sciences, than are those tasks required by the verbal and analytical tests. If this is the case, then the physical science majors would have received more experience in these types of items than would their companions in other fields of study who did not concentrate in those courses.

We shall shortly compare the performance of the various subgroups on an item-by-item basis for all the items that make up the verbal, quantitative and analytical sections of one form (Form 3FGR2) of the GRE test, searching for items which appear to favor certain subgroups differentially. Prior to that, we consider two additional analyses of the overall scores on the aptitude tests.

Analysis of the "Verbal" and "Mathematical" Fields of Study

We recognized that the four major fields of study chosen for the main analysis were necessarily broad and heterogeneous. Since our aim was to discern the potential impact of coursework on the verbal, quantitative and analytical aptitude test scores, we thought it useful to focus attention on students in subfields that were more clearly "verbal" or "mathematical" than others in the same subfields. Accordingly, a random subsample of about 40% of the total group was identified and only those of the four fields that were most clearly associated with verbal and mathematical course content were chosen--namely, humanities and physical science--ignoring, in this analysis, the "intermediate" fields of social science and biological science that are perhaps even more heterogenous and, presumably, less clearly associated with either verbal or mathematical.

Further selection was also undertaken. From the humanities group only those students were selected who had majored in particular subfields that were thought to capitalize on, or to develop, even more than the other subfields in the larger category of humanities, the verbal talents of the students; correspondingly, from the physical science group only those students were selected who had majored in particular subfields that were thought to capitalize on, or to develop, even more than the other subfields in the larger category of physical science, the mathematical talents of the student.[1] Other cases falling into these two major fields were abandoned for purposes of this substudy. The numbers of cases finally selected for the study of the "verbal" and "mathematical" fields are given in the following table:

Newly Constituted Major Fields	Men	Women	Totals
"Verbal" Humanities	257	1,577	1,834
"Mathematical" Physical Sciences	1,441	445	1,886
Totals	1,698	2,022	3,720

Using this "purified" subsample of students, which includes only those students in clearly verbal fields or clearly mathematical fields, we again developed prediction equations relating the student's aptitude test scores on each of the GRE-verbal, GRE-quantitative, and GRE-analytical tests to that student's aptitude test scores on both the SAT-verbal and SAT-mathematical tests. In order to allow the relationships to vary according to the student's sex and field of study, we fit a separate regression equation (for each of the three GRE scores) within each of the four sex-by-field-of-study subgroups. The results appear in Tables 12, 13, and 14. (To allow for comparisons with the predictions based on the unpurified sample, both SAT-

[1]The subfields retained for the study of "verbal" Humanities included: Classical Languages, Comparative Literature, English, Far Eastern Languages and Literature, French, German, Near Eastern Languages and Literature, Spanish, Other Foreign Languages, and Journalism. The subfields retained for the study of "mathematical" Physical Science included: Mathematics, Applied Mathematics, Statistics, Computer Sciences, Physics, and Chemical, Aeronautical, Civil, Electrical, and Mechanical Engineering.

verbal and SAT-mathematical scores were used in the prediction of GRE-analytical. In contrast, GRE-verbal was predicted by SAT-verbal only and GRE-quantitative was predicted by SAT-mathematical only, as was done for the unpurified sample.)

Table 12 shows the predicted values of GRE-verbal scores, by sex and field of study, for low verbal initial ability and high initial verbal ability students. As previously, we define low initial verbal ability students to have SAT-verbal scores of 380 (the 10th percentile of the total unpurified sample); the high initial ability students are defined to have SAT-verbal scores equal to 650 (the 90th percentile of the total unpurified sample). Upon comparing this table with the predictions based on the "unpurified" sample in Table 6, we see that the predicted GRE-verbal values for the "verbal" humanities students of both sexes and both levels of initial ability are greater than the matching predictions for the full sample of the humanities students. That is, students in the "verbal" humanities are predicted to do better on the GRE-verbal than students of matching initial abilities who are in the less homogeneously verbal subfields of the humanities. The gain is substantial for the lower ability men but small for men at the higher level of initial ability and for women at either level. In contrast, the predicted GRE-verbal scores for the "mathematical" physical science students of either level of initial ability are essentially the same as the corresponding scores for the full sample of physical science students. That is, students in the "mathematical" physical sciences, of either level of initial verbal ability and either sex are predicted to do essentially the same on the GRE-verbal as students of matching ability and sex who are in the less mathematical subfields of physical sciences.

As a result of the purification of the sample, we can see a moderate impact of field of study on scores on the GRE-verbal test with students in the "verbal" humanities having a consistent advantage over students in the "mathematical" physical sciences, particularly so for the students of lower initial ability. This effect of field of study on GRE-verbal scores was not apparent when the full (unpurified) sample was examined.

As we found with the unpurified data, the impact of field of study on GRE-quantitative scores is notably more pronounced than the effect of field of study on the verbal scores. The pertinent data are given in Table 13, which shows the predicted GRE-quantitative scores for students in the purified sample. The table shows that the students in the "mathematical" physical sciences have a strong and consistent advantage over their fellow students in the "verbal" humanities, especially so at the lower level of initial ability (initial ability levels as defined above). A comparison of Table 13 with Table 7, which provides the predictions for the unpurified sample, shows that the effect of purification is to enhance the measure of impact without changing any of the basic conclusions that were reached based on the unpurified data.

Finally, Table 14 shows the predicted values of GRE-analytical score for the students in the purified sample. Generally, the predicted scores for students in the "mathematical" physical sciences are slightly higher than the scores of the students in the "verbal" humanities. The exception to this generalization is that of the lower initial ability men, for whom the opposite is true. A comparison of this table with the corresponding results for the unpurified sample, shown in Table 11, shows that the predicted score of the lower ability men in the "verbal" humanities is noticeably higher than the corresponding scores for the unpurified sample. Apart from this, the scores for the purified sample are quite similar to those in the unpurified sample.

The final picture is that the conclusions reached from studying the purified sample are, in the main, unchanged from the conclusions obtained from the study of the full (unpurified) sample, although the effects are generally enhanced.

Analysis of the effects of college attended on aptitude test scores

In the last few sections we have considered the relationships between measures of "initial" academic aptitude, i.e., SAT scores at the onset of the college career, and measures of "final" academic aptitude, i.e., GRE scores at the close of undergraduate study. Our primary aim has been to determine if the relationships between initial and final measures of academic aptitude might be affected by the particular course of study selected. To this end, we have examined how the predicted scores on the GRE aptitude tests, for given scores on the SAT tests, vary by field of study.

These analyses, however, while considering a student's sex and field of study in forming the predictions, do not consider another potentially important characteristic of the student: the college attended. Students in different colleges may have studied differently constituted coursework even while majoring within the same subfields of humanities, social science, biological science, or physical science. It is plausible that the particular pattern of coursework, as well as the academic environment, would affect the final outcome measure, namely the scores on the GRE aptitude tests, for a student considered in this study. Moreover, it is possible that the male and female students in the sample coming from the different fields of study may have been drawn from their colleges in disproportionate numbers--for example, larger numbers of students of one sex in social science (say), coming from lower-scoring colleges with less demanding courses, and larger numbers of students of the other sex in physical science (say), coming from higher-scoring colleges with more demanding courses. Accordingly, our interest in the present analysis lies in ascertaining the extent to which the relationship between initial ability and final ability varies by institution attended.

The obvious way to address this issue would be to compute a separate set of prediction equations for each of the institutions represented in our dataset, basing the predictions for a given institution only on the students who have attended that institution, and then comparing the resulting within-institution prediction equations. Unfortunately, this direct approach is not applicable in our situation. Of the institutions represented in our study, 73% had fewer than 10 students taking one of the two forms of the GRE, with an average of three students per institution. Under such circumstances the estimates of the parameters defining the within-school prediction equations for those schools would be, at best, seriously unstable because of the small number of students within each school available for calculating those prediction equations.

A successful approach to solve the estimation problems associated with small within-institution sample sizes is to employ an analysis technique which takes the hierarchical nature of the data (students within institutions) into account, such as the variance component modelling advocated by Aitkin and Longford (1986). In variance component analysis we focus our attention on estimating how much of the total variation in individual scores may be attributed to characteristics of the individual students and how much may be attributed to the environment in which the individual students are placed. That is, we will want to partition the total variation of scores into two types of components: those reflecting variation among students within institution and those reflecting variation among institutions. The magnitudes of the between-institution components, relative to those of the within-institution components, would then be indicative of the importance of the institution attended on the outcome measure (scores on the GRE test), relative to the importance of individual characteristics (within schools) on the outcome.

Because the impact of curriculum was relatively high for measured quantitative aptitude (GRE-quantitative scores) and low for the other two measures of outcome academic aptitude (GRE-verbal and GRE-analytic scores), it is likely that the effect of institution attended will be the greatest for the measure of quantitative aptitude. Consequently, we decided to ascertain the importance of institution attended only on the score on the GRE-quantitative test. We base this analysis on the quantitative scores of the respondents to form K-3FGR3 of the GRE. The fitting of variance component models to this data was carried out with the VARCL program of Longford (1986). This program employs a Fisher scoring algorithm to provide maximum likelihood estimates of all regression coefficients and variance components.

To improve the stability of the estimates, the dataset was restricted to those institutions with at least 10 students responding to the form K-3FGR3. The redefinition of the sample in this fashion resulted in a total of 7954 students from 292 institutions, with an average of 27 students per institution. (The full set of 10,322 respondents to Form K-3FGR3 came from 1,067 institutions; the average number of students per institution in the 775 excluded institutions was 3).

It is interesting to note, in passing, that one effect of restricting the dataset to the schools with at least 10 respondents to the form is to raise the means of both the SAT-mathematical and the GRE-quantitative scores by 17 and 18 points, respectively. (The SAT-mathematical and GRE-quantative mean scores for the full set of 10,322 respondents were 556 and 573. The corresponding mean scores for the restricted set of 7954 were 573 and 591).

A series of variance component models were fit to this data. The results of one model fit, which was selected as providing an adequate description of the data, are shown in Table 15. The top portion of the table provides the estimates, along with standard errors, of the fixed-effect parameters for the following variables:

Field of Study
Sex
SAT-mathematical score = SAT-M
A quadratic transformation of SAT-M: $SATMSQ = (SATM - 500)^2/100$
Interactions of SATM and SATMSQ with major

(The quadratic transformation of SAT-M, SATMSQ, was included to capture the slight curvilinearity of the relationship between the SAT and the GRE scores. The predictor was centered by subtracting 500 and scaled by 100 to enhance the numeric stability of the estimates.) Using these fixed-effect parameter estimates produces predicted scores on the GRE-quantitative test which are, on the whole, consistent with previously reported predictions (e.g. Table 7).

The random portion of the model includes terms addressing the following components of variance:

σ^2_I : the between-institution variance in intercept;

σ^2_S : the between-institution variance in the slope of the line relating the SAT-mathematical score with the GRE-quantitative score;

σ^2_ε : the student-level variance.

The estimates of these variance components are provided in the lower part of Table 15. Also included are the square roots of the variance components ("sigma") and, for the institution level components, estimates of the standard errors of the sigmas. (The magnitudes of other components were too small to deserve mention.)

The magnitudes of the estimated variance components for the institution-level random parameters indicate that both the intercept and the slope of the regression line relating GRE-quantitative score (adjusted for major and sex) on SAT-mathematical score vary across institutions in a statistically significant manner. Of interest is the importance of this variation across institutions from a practical viewpoint.

We can address this question by considering how much of the total variability in a predicted score is attributable to institutional-level variance and how much to student-level variance. Since the variance component model includes a random slope (on SAT-M), this partitioning of variance must depend on the level of initial ability, the SAT-mathematical score.

The estimated total variance of a predicted (adjusted) GRE-quantitative score, given a SAT-mathematical score of X, is

$$\hat{\sigma}^2_{T(X)} = \hat{\sigma}^2_{\varepsilon} + \hat{\sigma}^2_{I} + 2X\hat{\sigma}_{IS} + X^2\hat{\sigma}^2_{S} ,$$

where $\hat{\sigma}_{IS}$ is the estimated covariance between the random intercept and the random slope. Using the values from Table 15, we can calculate the proportions of the total variance attributable to student-level variance as $\hat{\sigma}^2_{\varepsilon}/\hat{\sigma}^2_{T(X)}$. The table below shows the result for selected values of SAT-mathematical score. (The values for 573 are shown because 573 is the mean SAT-mathematical score for the set of data under current analysis.)

SAT-M	Total Variance	Percent of Total Variance Attributable to Student Level
350	3227.4	99.93%
573	3309.9	97.44%
750	3460.0	93.21%

Consequently, for SAT-mathematical scores in the range of 350 to 750, at most seven percent of the total variability of a predicted GRE-quantitative score is attributable to institutional-level variance and no more than three percent of the total variability is institutional-level for SAT scores below the mean value of 573. The interpretation of this result is that, for a given level of initial ability, the variation in GRE-quantitative scores between students within an institution swamps the variation in scores between institutions. This might be taken to indicate that, given initial ability, individual (i.e., within-school) characteristics are much more important in determining the final GRE score than are institutional level characteristics. We do, however, note that the percent of the total variability in predicted scores attributable to institutional-level variables increases as the level of the initial ability increases. That is, there is more between-institution variability in predicted scores for the higher initial ability students so that institutional level characteristics are relatively more important for the higher ability students. This might indicate that certain schools are more effective instructionally than others for the higher ability students.

Analysis of differential item performance by sex and field of study

We now turn to the final phase of our study which is to study the items that make up the GRE-verbal, GRE-quantitative and GRE-aptitude tests. In this study we seek to identify and understand, if possible, the mechanisms in those aptitude items that are relatively more, and those that are relatively less, resistant to general educational experiences in college. In particular, we seek to heighten our understanding of items, and item types, that are more likely to reflect, or, conversely, to resist the effects of curriculum or sex. We specifically wish to discover if students oriented differently with respect to the various fields of study and/or of different sex will have the same degrees of success on the various items of a GRE general test, after controlling for initial ability as measured by their previous SAT-verbal and SAT-mathematical Test scores. We will approach this question by applying the Mantel-Haenszel procedure (Mantel and Haenszel, 1959; Holland and Thayer, 1986) to the responses to the items of one of the two forms that we have been studying: Form 3FGR2.

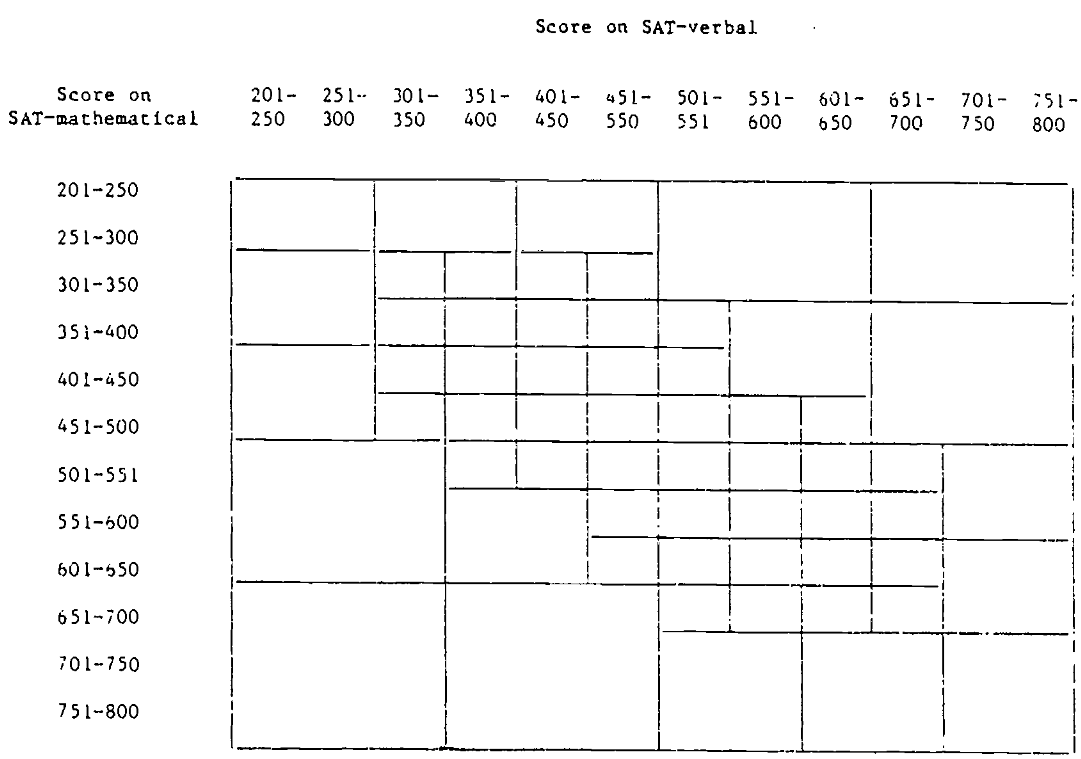
The Mantel-Haenszel procedure compares the performance of two groups of examinees, called the focal group and the reference group, on an item-by-item basis, providing for each item a measure of the differential item performance of the focal group as compared to the reference group. This measure will be called MH D-DIF, where DIF is an acronym for <u>differential item functioning</u>

and the D indicates that the measure of DIF is roughly on the ETS delta scale.

Seven parallel Mantel-Haenszel analyses of the responses to the items in the Form 3FGR2 of the GRE General Aptitude Test were conducted. In each of these analyses the male students in the humanities served as the reference group. Each of the 7 remaining sex-by-field-of-study subgroups was selected in turn as the focal group for comparison with the common reference group.

Although the choice of the reference group is necessarily arbitrary, the male humanities group was chosen as the reference group because the members of this group are at one extreme of the conceptual continuum between heavily verbal and heavily mathematical fields of study. It is likely that using such a group as the common reference group will enhance the detection of differential item functioning due to field of study, particularly for the comparisons of the two extreme groups: humanities vs. physical science.

Each of the 7 Mantel-Haenszel analyses involved the comparison of the designated focal group (say the men in the physical sciences) with the reference group in terms of their performance on each of the 186 items which made up the Form 3FGR2 of the GRE General Aptitude Test, a comparison that was done one item at a time. In conducting a comparison of the two groups in terms of their responses to a given item (the studied item) we are seeking indications of differential item functioning (DIF), by which we mean differences in the performance on the studied item between members of the focal group and reference group who are of comparable initial abilities. Our first step in the analysis of DIF for a given focal group was therefore to match the members of the focal group with members of the reference group on the basis of their scores on SAT-verbal and SAT-mathematical. The matching was accomplished by first classifying each student into one of 60 categories, based on that student's SAT-verbal and SAT-mathematical score, then using these categories to match the members of the focal group with the members of the reference group. The categories for matching, indicated in the schematic diagram shown below, were devised to allow a reasonably close match of students in terms of their scores on both tests while ensuring that every cell contained students in both the focal group and the reference group. Note that the cells increase in size as a function of their distance from the center of the bivariate diagram in order to accomodate the smaller numbers of cases in the extreme intervals. These larger cells correspond to students with (typically) very high levels of ability on one SAT test and (typically) very low levels of ability on the other SAT test.



After matching on the basis of initial ability, we compare the performance on the studied item for each matched category of students in the focal group and the reference group. For the j^{th} matched category, let $n^{R}j$ be the number of students who are also in the reference group and let $p^{R}j$ be the proportion of these who responded correctly to the studied item. Similarly, define $n^{F}j$ and $p^{F}j$ for the focal group students who are in the j^{th} matched category. Define the odds ratio α_j by:

$$\alpha_j = \frac{p^{R}j}{q^{R}j} \Big/ \frac{p^{F}j}{q^{F}j} ,$$

where $q^{R}j = 1 - p^{R}j$ and $q^{F}j = 1 - p^{F}j$. If there is no difference in the performance on the studied item of the members of the focal and reference groups within the j^{th} match set then α_j will be equal to 1. Otherwise, the item is functioning differentially within the j^{th} matched category--in favor of the reference group if $\alpha_j > 1$ and in favor of the focal group if $\alpha_j < 1$.

The Mantel-Haenszel procedure estimates that common odds-ratio, α, across all of the 60 matched categories. This common odds ratio is estimated by

$$\hat{\alpha} = \frac{\Sigma\, p^{R}j\; q^{F}j\; n^{R}j\; n^{F}j \;/\; (n^{R}j + n^{F}j)}{\Sigma\, q^{R}j\; p^{F}j\; n^{R}j\; n^{F}j \;/\; (n^{R}j + n^{F}j)} ,$$

and is the average factor by which the odds that a member of the reference group will respond correctly to the item exceeds the odds for a comparable member of the focal group. Values of α accordingly provide a measure of the amount of differential item functioning.

More conveniently, we will use MH D-DIF = $-2.35 \ln(\hat{\alpha})$ as our measure of differential item functioning. This transformation centers the measure of DIF about the value 0, corresponding to the absence of differential functioning. The multiplier -2.35 puts the measure on a scale comparable to the ETS "delta scale" and <u>reverses</u> the measure so that positive values indicate DIF in favor of the focal group and so that negative values indicate DIF in favor of the reference group. There is a convenient approximate linear relation between the values of MH D-DIF and the difference in the values of the proportions of correct responses between matched members of the focal group and the reference group for items of moderate difficulty (in the range of 30% to 70% correct responses). The absolute value of MH D-DIF is roughly equal to 10 times the absolute difference in the proportions of correct responses between the focal group and the matched members of the reference group. Thus |MH D-DIF|=1 corresponds to a difference of 10 percentage points and |MH D-DIF|=2 to a difference of 20 percentage points.

The results of the 7 Mantel-Haenszel analyses of the Form 3FGR2 of the GRE General Test are shown in Tables 16, 17, and 18. Table 16 shows characteristics, by focal group, of the distributions of the values of MH D-DIF across the 76 items which constituted the verbal sections of the test. Included in the table are selected order statistics from these distributions (the minimum, maximum, median and quartiles) as well as the average value of MH D-DIF and two measures of the variability (mean-squared errors) of the DIF statistics. The first measure of the variability of the MH D-DIF statistics is the within-item variance--essentially, the stability of the MH statistic--computed as the average of the estimated variances of the 76 DIF statistics where the estimate of the variance of a given MH D-DIF statistic is based on the approximation of Holland and Phillips (1987). The second measure of the variability of the distribution of the MH D-DIF statistics is the between-item variance of the estimates of DIF, the measure of the variability from item to item of the D-DIF statistics.

In addition to showing the described statistics--in which the male humanities group is taken as the reference group and each of the other 7 as the focal group--Table 16 shows the same statistics, by item, for the average of the male and female values of DIF within the social science, biological science, and physical science fields of study. Equivalent statistics also appear for the difference by item between the male and female MH D-DIF values, again by field of study. The difference in values of MH D-DIF between men and women in a given field of study is an estimate of the value of DIF that would have been obtained if the women in that field of study were compared, as the focal group, to the men in the field of study as reference group.

Tables 17 and 18, respectively, impart the same information for the distributions of MH D-DIF statistics across items in the quantitative and the analytical sections of the test.

On examining these tables we see that, for the verbal items (Table 16), the average values of the DIF statistics are generally relatively small with, for example, the most extreme mean value being −.43, disfavoring the female physical science students. This corresponds to an average difference in percent correct on the items between members of this focal group and their matched cohorts in the reference group of around 4%. The other statistics in this table, apart from the extremes, indicate effects of differential item functioning generally less than this magnitude.

The quantitative items, on the other hand, appear to show a greater indication of differential item functioning. (See Table 17.) This is particularly so for the students in the physical sciences where the mean and median values of the DIF statistics are greater than 1, thus favoring the students in the physical sciences. Such a result should not be surprising since the physical science students had an advantage (in terms of predicted score) on the quantitative test, even after controlling on SAT scores.

Also in line with previous results are the summary statistics for the analytical test (Table 18), which are, on the whole, moderate in magnitude--showing larger values of D-DIF than those of the verbal test, but smaller than those of the quantitative test.

As another view of the distributions of the MH D-DIF statistics we present Tables 19, 20 and 21, for the verbal, quantitative and analytical tests, respectively. In each of these tables we have classified the items, for each focal group, ccording to the value of their MH D-DIF into three classes: relatively minor effects (|MH D-DIF|<1), moderate effects 1≤|MH D-DIF|<2), and strong effects (|MH D-DIF|≥2). We have further classed the items with more than minor DIF effects by the group favored (focal or reference).

We see again from Table 19 that there is a relatively small amount of differential item functioning across the subgroups for the verbal items, with most items being classed into the minor effect category. There is some indication of item functioning which differentially disfavors the students in the physical sciences. Few items were classed as being of strong DIF effects.

The data for the quantitative items, shown in Table 20, paint a different picture of differential functioning. In this case more than half of the items are classed as at least moderately favoring the physical science students and about half of those quite markedly so. There is also an indication that the male biology students fare somewhat better than the matched male humanity students on the quantitative items.

The indications from Table 21 are that, on the whole, the extent of differential item functioning for the analytical items is generally minor.

It was also thought desirable to have an omnibus measure of the degree of differential functioning exhibited by a given item across all eight of the sex-by-field-of-study subgroups. In constructing this measure, we will consider the distributions of the sizes of MH D-DIF relative to their estimated standard errors. Let

$$Z_i = (\text{MH D-DIF}_i)/(SE_i),$$

where MH D-DIF_i is the value of the DIF estimator for one of the seven focal groups relative to the reference group (the male humanities students) and SE_i is its standard error (the square root of the Holland-Phillips variance estimate). Assuming that there is no differential item functioning for any of the focal groups relative to the common reference group, each of the statistics $Z_1, \ldots, Z_7$ will asymptotically have a standard normal distribution. Since each of the seven groups has been compared to a common reference group, the Z's are positively correlated. Under the additional assumption that the correlation between each pair Z_i and Z_j of the Z's is constant and equal to $\frac{1}{2}$, it can be shown that the corrected sum of squares,

$$S = 2(\Sigma Z_j^2 - \frac{1}{8}(\Sigma Z_j^2),$$

is approximately distributed like a chi-squared random variable with 7 degrees of freedom. (Empirical evidence based on the data in hand suggest that the main part of the distribution of S is, in fact, well approximated by a chi-squared distribution but that the degrees of freedom are between 4 and 5. This reduction in the degrees of freedom is due in part to the fact that the values of DIF for the items are never exactly zero.)

We will use S as our overall measure of differential functioning of an item across the eight sex-by-field-of-study subgroups interpreting large values, relative to what we would expect given the assumed distribution, as indications of differential item functioning for at least one of the subgroups in question.

Table 22 shows the numbers of items in each of the verbal, quantitative and aptitude sections of the test whose value of the statistic S (as defined above) exceeds certain selected critical values corresponding to the 90th, 95th and 99th percentiles from the chi-squared distribution with 7 degrees of freedom. The number of quantitative items, 11, whose statistic S exceeds the 99th percentile is striking, inasmuch as it is many times more than would be expected under the null assumption. (This number would be even more striking if the degrees of freedom were lower, say 5). Of these 11 extreme items, the large majority--9 items--favor the physical science students (both sexes roughly equally favored over the male humanities students). The numbers of extreme items from the verbal and analytical sections of the form are closer to those expected under the assumptions. Of the three most extreme verbal items, two apparently disfavor all women; the other favors biology students of either sex. Only one analytical item appears extreme at any reasonable level and that favors men in the biological sciences. These results indicate that the impact of field of study is the greatest for the quantitative items, but is not of great import for the verbal or analytical items.

There was an effort made at the item level to try to correlate, at an "eyeball" level, the Mantel-Haenszel values found in the study with the particular content of the items themselves. This process yielded no success for the mathematical and analytical items. Although the items did differ from one another with respect to impact (after controlling on SAT scores), they did not fall into identifiable categories that would make it possible to predict which items would be likely to show such impact and which would not. There was a hint, however, that verbal items that made reference to language or concepts that were well known to one subgroup but not necessarily to others might be giving that subgroup a slight advantage. However, there were very few such items and the authors are reluctant to draw strong conclusions on the basis of such weak observations.

SUMMARY AND DISCUSSION

A sample of 22,923 students who had taken the GRE General Test in the academic years 1983-84 and 1984-85 and who had also taken the SAT four or five years earlier were identified, and classified by undergraduate field of study (four major categories of curriculum) and sex. Several analyses were undertaken to determine the differential impact that undergraduate field of study might exert on GRE-verbal, GRE-quantitative, and GRE-analytical scores, after controlling on SAT-verbal and mathematical scores; and to determine if that impact varied by sex. It was found that the correlations of SAT-verbal with GRE-verbal and SAT-mathematical with GRE-quantitative were extremely high; both correlations were .86 across the entire sample, and ranged in the low to middle .80's in the eight sex-by-field-of-study subgroups. The impact of curriculum and sex was found to be low on GRE-verbal and GRE-analytical scores, but relatively high for GRE-quantitative. Further studies designed to "purify" the fields of study and include only clearly verbal fields and clearly mathematical fields--omitting entirely students in social and biological science--showed enhanced impact, but not of great magnitude. An additional study indicated that, after accounting for major field of study and initial ability, the effect of the institution attended on GRE-quantitative score is generally slight, although the importance of institution is greater for the higher ability students. Although the additional studies helped a bit to clarify the results, the basic conclusions, that the curricular impact on GRE scores was quite small for verbal, slightly greater for analytical, but substantial for quantitative, remain unchanged.

Although it was expected that both the actual and predicted means on the quantitative tests would be higher for the men than for the women, we found that the means on the verbal tests both within field of study and across the entire sample were, with one exception (in the humanities), higher for men than for women. What is more surprising, even when conditioned on SAT-verbal scores, the GRE-verbal means for the men were higher than those for the women. This was found to be the case without exception in all four fields of study and at both high and low levels of ability. The reverse, however, was true for the GRE-analytical test; although the means for the men were consistently higher than those for the women, we found that when the students were conditioned on both SAT-verbal and SAT-mathematical, the women's means on GRE-analytical consistently exceeded those of the men.

In a separate phase of the study an attempt 'as made to identify the kinds of items that were relatively resistant to curricular and sex effects. These analyses measured differential item functioning by use of the Mantel-Haenszel technique, in which the odds of the students answering each GRE item correctly were compared across groups, after matching on both SAT-verbal and -mathematical scores. In these analyses the male humanities group was taken as the "reference" group, and each of the other 7 groups was individually taken as the "focal" group and compared with it. It was found that the items in the GRE-analytical section showed the smallest proportion of significant DIF (differential item functioning) values. The items in the GRE-verbal section showed a somewhat greater proportion of significant DIF values, and the GRE-quantitative section showed the largest proportion of such items. It is surmised that exposure in college to the physical sciences and their mathematical content does have a disporportionate effect on later GRE scores. This effect, however, appears to apply minimally to the GRE-verbal and only slightly more to GRE-analytical items.

An attempt made to correlate informally the DIF values found in this study with the content of the items yielded no success for the mathematical and analytical items. There was a hint, however, that verbal items that made reference to language or concepts well known to one subgroup but not to another might be giving that subgroup some advantage. However, there were very few such items in the verbal test, and therefore this conclusion can only be considered tentative.

The results of this study have confirmed what we have already known, or suspected, about the differential impact of educational experiences in late adolescence on aptitude test scores. At least at this level of age and education, it matters relatively little whether a student concentrates his or her studies in the humanities, social studies, biological sciences, or physical sciences or whether the student is male or female; scores on the verbal section of the GRE General Test are very much the same regardless of field of concentration or sex. This is not so true for the GRE-quantitative Test. There, it appears, students of the same ability level, as measured by the SAT, but who have spent their undergraduate years in the study of mathematics or mathematically-related subjects do better on the GRE-quantitative Test than those who have not; and the more concentrated the study or use of undergraduate mathematics, the higher the quantitative score. It also appears that women of the same initial ability as men (again, as measured by the SAT) who have studied the same general curriculum in college earn somewhat lower scores on the GRE-quantitative test (although the difference vanishes for students of high initial ability in the clearly verbal or clearly mathematical fields of study.) The differential effect on the analytical score is intermediate between that of the verbal score and that of the quantitative score. This latter outcome is not overly surprising since this study, and other studies (e.g., Angoff & Cowell, 1986) conducted earlier, have found that the GRE-analytical Test correlates substantially with both GRE-verbal and GRE-quantitative--somewhat higher with quantitative than with verbal.

It is interesting to speculate on the reasons for the difference in curricular impact on the quantitative score. Perhaps the principal reason is that the GRE-quantitative Test is more nearly an achievement test in the usual sense, consisting specifically of content learned in the early secondary school curriculum at a time not much more than seven years before the student takes the GRE. Inasmuch as achievement test material is by definition highly susceptible to educational intervention, it is not at all surprising that students who have used their mathematics during their normal work in college would have honed their understanding and skills on this material, and generally in proportion to their use of it. It is also plausible that students who have not used their mathematics in recent years may have lost some of their earlier understanding and skills, and generally in proportion to their lack of use of them.

It would therefore appear that if we were to search for quantitative items that would be less subject to the effect of study than those found in the GRE General Test, we would have to select them from content areas that are as far removed as possible fro the formal mathematics learned in school, in recent years, at least. Indeed, it would be desirable to see to it that they contain no more formal mathematics than that learned in the very early grades of school--involving no more than the four basic arithmetical operations, if possible.

While the foregoing approach to mathematical aptitude might be expected to introduce greater resistance to curricular effects than is true of the type of test in use today, a reasonable conjecture is that even this approach will not result in the kind of stability characteristic of verbal aptitude. Casual observation suggests that individuals with advanced or concentrated mathematical training seem to possess mathematical insights in solving difficult problems, even those that call on the use of no more than elementary operations known to everyone, insights that other individuals without such training do not have.

In any case, even with the instruments currently available, it appears that the usual experience of pursuing a particular course of study in college has little effect on verbal or analytical aptitude, but a substantial effect on quantitative aptitude, at least as measured by the GRE General Test. The correlation of SAT-mathematical with its counterpart, GRE-quantitative, is in the middle .80s, as is the correlation of SAT-verbal with GRE-verbal. The fact that the former correlation is as high as it is in spite of the greater impact of curriculum on quantitative aptitude is explainable by the observation that the variation in the mathematical-quantitative surface is greater than that in the verbal-verbal surface in both the diagonal <u>and</u> the off-diagonal directions.

REFERENCES

Aitkin, M. & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. Journal of the Rcyal Statistical Society, A, 149, 1-43.

Anastasi, A. (1975). Harassing a dead horse. (Review of D. R. Green (Ed.) (1974). The aptitude-achievement distinction. Monterey, CA: CTB/McGraw Hill.) The Review of Education, 1, 356-362.

Anastasi, A. (1980). Abilities and the measurement of achievement. In W. B. Schrader (Ed.) New directions for testing and measurement. San Francisco: Jossey-Boss. Pp. 1-10.

Anastasi, A. (1984). Influences on aptitude and achievement test development and usage. In B. S. Plake (Ed.) Social and technical issues in testing: Implications for test construction and usage. Hillsdale, NJ: Lawrence Erlbaum Associates. Pp. 129-140.

Angoff, W. H. & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. Journal of Educational Measurement, 23, 327-345.

Bereiter, C. (1974). Comments on Burket's paper: Eliminating an inconsistency in Burket's time-distance model. In D. R. Green (Ed.) The aptitude-achievement distinction. Monterey, CA: CTB/McGraw-Hill. Pp. 49-53.

Carroll, J. B. (1974). The aptitude-achievement distinction: The case of foreign language aptitude and proficiency. In D. R. Green (Ed.) The aptitude-achievement distinction. Monterey. CA: CTB

REFERENCES (cont'd.)

Donlon, T. F. (1984). (Ed.) The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests. New York: The College Board.

Educational Testing Service (1984). GRE 1984-85 Information Bulletin. Princeton, NJ: Author.

Educational Testing Service (April, 1985). Graduate Record Examinations General Test. Test analysis report for Forms 3FGR2 and S-3FGR2. SR-85-50. Princeton, NJ: Author.

Educational Testing Service (June, 1985). Graduate Record Examinations General Test. Test analysis report for Forms 3FGR3 and S-3FGR3. SR-85-90. Princeton, NJ: Author.

Educational Testing Service (October, 1987). Summary of Admission Testing Program scores, 1965-66 through 1986-87. Princeton, NJ: Author

Educational Testing Service (1987). Guide to the use of the Graduate Record Examinations Program, 1987-88. Princeton, NJ: Author.

Green, B. F. (1978). In defense of measurement. American Psychologist, 33, 664-670.

Holland, P. W. & Phillips, A. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. Biometrics, 43, 425-432.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.) Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

REFERENCES (cont'd.)

Humphreys, L. G. (1974). The misleading distinction between aptitude and achievement tests. In D. R. Green (Ed.) The aptitude-achievement distinction. Monterey, CA: CTB/McGraw-Hill. Pp. 262-274.

Longford, N. (1986). VARCL - Interactive software for variance component analysis. The Professional Statistician, 5, 28-32.

Mantel, N. & Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Table 1

Numbers of Students in the Entire Study Sample,

Classified by Ethnic Group, Field of Study, and Sex

Ethnic Group	Humanities		Social Science		Biological Science		Physical Science		Total	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Black[a]	29	86	165	454	43	135	126	106	363	781
Hispanic[b]	31	39	90	153	33	58	81	24	235	274
Asian[c]	25	51	58	124	60	79	232	85	375	339
White	1,163	1,883	2,576	4,586	1,354	2,606	3,977	1,494	9,070	10,569
Other[d]	25	33	70	87	25	46	90	24	210	190
Missing[e]	32	49	72	110	46	45	120	43	270	247
Total	1,305	2,141	3,031	5,514	1,561	2,969	4,626	1,776	10,523	12,400

These ethnic groups are defined as follows:

[a]Black or Afro-American
[b]Mexican American or Chicano, Puerto Rican, and other Hispanic or Latin American
[c]Oriental or Asian American
[d]American Indian, Eskimo or Aleut, and others not otherwise defined
[e]Ethnic group not available

Table 2a

Intercorrelations Between SAT and GRE Scores

for the Total Study Sample

N = 22,923

	SAT Verbal	SAT Mathe-matical	GRE Verbal	GRE Quanti-tative	GRE Analyt-ical	Mean	Standard Deviation
SAT Verbal	1.000	.628	.858	.547	.637	518.8	104.7
SAT Mathematical	.628	1.000	.598	.862	.734	556.0	110.2
GRE Verbal	.858	.598	1.000	.560	.649	510.1	107.7
GRE Quantitative	.547	.862	.560	1.000	.730	573.4	125.6
GRE Analytical	.637	.734	.649	.730	1.000	579.7	117.6

Table 2b

Intercorrelations Between SAT and GRE Scores

for the Male Humanities Sample

N = 1,305

	SAT Verbal	SAT Mathematical	GRE Verbal	GRE Quantitative	GRE Analytical	Mean	Standard Deviation
SAT Verbal	1.000	.629	.868	.558	.633	548.0	105.5
SAT Mathematical	.629	1.000	.594	.857	.714	574.6	105.3
GRE Verbal	.868	.594	1.000	.563	.641	547.7	115.5
GRE Quantitative	.558	.857	.563	1.000	.722	575.2	113.9
GRE Analytical	.633	.714	.641	.722	1.000	579.8	117.6

Table 2c

Intercorrelations Between SAT and GRE Scores

for the Female Humanities Sample

N = 2,141

	SAT Verbal	SAT Mathematical	GRE Verbal	GRE Quantitative	GRE Analytical	Mean	Standard Deviation
SAT Verbal	1.000	.599	.866	.544	.600	551.2	102.6
SAT Mathematical	.599	1.000	.577	.834	.684	532.2	98.3
GRE Verbal	.866	.577	1.000	.550	.620	535.2	110.4
GRE Quantitative	.544	.834	.550	1.000	.698	517.4	111.2
GRE Analytical	.600	.684	.620	.698	1.000	568.2	110.7

Table 2d

Intercorrelations Between SAT and GRE Scores

for the Male Social Science Sample

N = 3,031

	SAT Verbal	SAT Mathematical	GRE Verbal	GRE Quantitative	GRE Analytical	Mean	Standard Deviation
SAT Verbal	1.000	.655	.857	.591	.636	510.3	106.0
SAT Mathematical	.655	1.000	.619	.846	.735	541.0	106.7
GRE Verbal	.857	.619	1.000	.603	.649	513.4	111.6
GRE Quantitative	.591	.846	.603	1.000	.748	557.4	117.5
GRE Analytical	.636	.735	.649	.748	1.000	561.5	119.6

Table 2e

Intercorrelations Between SAT and GRE Scores

for the Female Social Science Sample

N = 5,514

	SAT Verbal	SAT Mathe-matical	GRE Verbal	GRE Quanti-tative	GRE Analyt-ical	Mean	Standard Deviation
SAT Verbal	1.000	.658	.862	.590	.658	493.0	105.9
SAT Mathematical	.658	1.000	.632	.826	.715	499.6	98.8
GRE Verbal	.862	.632	1.000	.618	.678	481.2	105.2
GRE Quantitative	.590	.826	.618	1.000	.725	499.1	111.1
GRE Analytical	.658	.715	.678	.725	1.000	542.3	115.1

Table 2f

Intercorrelations Between SAT and GRE Scores

for the Male Biological Science Sample

N = 1,561

	SAT Verbal	SAT Mathematical	GRE Verbal	GRE Quantitative	GRE Analytical	Mean	Standard Deviation
SAT Verbal	1.000	.615	.832	.552	.599	502.8	97.3
SAT Mathematical	.615	1.000	.585	.814	.691	560.5	95.3
GRE Verbal	.832	.585	1.000	.579	.620	505.0	97.4
GRE Quantitative	.552	.814	.579	1.000	.712	598.3	100.1
GRE Analytical	.599	.691	.620	.712	1.000	579.3	109.5

Table 2g

Intercorrelations Between SAT and GRE Scores

for the Female Biological Science Sample

N = 2,969

	SAT Verbal	SAT Mathematical	GRE Verbal	GRE Quantitative	GRE Analytical	Mean	Standard Deviation
SAT Verbal	1.000	.659	.833	.627	.647	495.4	98.4
SAT Mathematical	.659	1.000	.623	.825	.715	521.1	96.5
GRE Verbal	.833	.623	1.000	.637	.660	482.6	96.6
GRE Quantitative	.627	.825	.637	1.000	.729	535.0	108.0
GRE Analytical	.647	.715	.660	.729	1.000	562.3	109.8

Table 2h

Intercorrelations Between SAT and GRE Scores

for the Male Physical Science Sample

N = 4,626

	SAT Verbal	SAT Mathematical	GRE Verbal	GRE Quantitative	GRE Analytical	Mean	Standard Deviation
SAT Verbal	1.000	.611	.844	.537	.605	543.6	97.5
SAT Mathematical	.611	1.000	.557	.794	.696	640.3	89.7
GRE Verbal	.844	.557	1.000	.537	.622	534.8	102.7
GRE Quantitative	.537	.794	.537	1.000	.698	686.5	83.0
GRE Analytical	.605	.696	.622	.698	1.000	633.5	107.6

Table 2i

Intercorrelations Between SAT and GRE Scores

for the Female Physical Science Sample

N = 1,776

	SAT Verbal	SAT Mathe- matical	GRE Verbal	GRE Quanti- tative	GRE Analyt- ical	Mean	Standard Deviation
SAT Verbal	1.000	.603	.858	.549	.613	541.9	99.8
SAT Mathematical	.603	1.000	.588	.813	.689	606.0	91.1
GRE Verbal	.858	.588	1.000	.581	.640	522.0	103.1
GRE Quantitative	.549	.813	.581	1.000	.722	645.2	91.3
GRE Analytical	.613	.689	.640	.722	1.000	630.1	108.0

Table 2j

Summary of Descriptive Statistics
for the Subgroups of the Study Sample

	Humanities		Social Science		Biological Science		Physical Science		Total Study Sample
	Men	Women	Men	Women	Men	Women	Men	Women	
No. of Cases	1,305	2,141	3,031	5,514	1,561	2,969	4,626	1,776	22,923
SAT-verbal									
Mean	548.0	551.2	510.3	493.0	502.8	495.4	543.6	541.9	518.8
Std. Dev.	105.5	102.6	106.0	105.9	97.3	98.4	97.5	99.8	104.7
SAT-mathematical									
Mean	574.6	532.2	541.0	499.6	560.5	521.1	640.3	606.0	556.0
Std. Dev.	105.3	98.3	106.7	98.8	95.3	96.5	89.7	91.1	110.2
GRE-verbal									
Mean	547.7	535.2	513.4	481.2	505.0	482.6	534.8	522.0	510.1
Std. Dev.	115.5	110.4	111.6	105.2	97.4	96.6	102.7	103.1	107.7
GRE-quantitative									
Mean	575.2	517.4	557.4	499.1	598.3	535.0	686.5	645.2	573.4
Std. Dev.	113.9	111.2	117.5	111.1	100.1	108.0	83.0	91.3	125.6
GRE-analytical									
Mean	579.8	568.2	561.5	542.3	579.3	562.3	633.5	630.1	579.7
Std. Dev.	117.6	110.7	119.6	115.1	109.5	109.8	107.6	108.0	117.6

Table 3a

Mean SAT-verbal Scores by Sex Within Field of Study with Across-Sex Average Scores and Male-Female Differences Within Field of Study (Standard Errors in Parentheses)

	Humanities	Social Science	Biological Science	Physical Science
Men	548.0(2.9)	510.3(1.9)	502.8(2.5)	543.6(1.4)
Women	551.2(2.2)	493.0(1.4)	495.4(1.8)	541.9(2.4)
M-W Average	549.6(1.8)	501.6(1.2)	499.1(1.5)	542.8(1.4)
M-W Difference	-3.2(3.6)	17.3(2.4)	7.4(3.1)	1.7(2.8)

One-way analysis of variance: $F = 181$; $ETA^2 = .052$

Table 3b

Mean SAT-mathematical Scores by Sex Within Field of Study with Across-Sex Average Scores and Male-Female Differences Within Field of Study (Standard Errors in Parentheses)

	Humanities	Social Science	Biological Science	Physical Science
Men	574.6(2.9)	510.3(1.9)	560.5(2.4)	640.3(1.3)
Women	532.2(2.1)	499.6(1.3)	521.1(1.8)	606.0(2.2)
M-W Average	553.4(1.8)	520.3(1.2)	540.8(1.5)	623.2(1.3)
M-W Difference	42.4(3.6)	41.4(2.3)	39.4(3.0)	34.3(2.6)

One-way analysis of variance: F = 908; ETA^2 = .217

Table 3c

Mean GRE-verbal Scores by Sex Within Field of Study with Across-Sex Average Scores and Male-Female Differences Within Field of Study (Standard Errors in Parentheses)

	Humanities	Social Science	Biological Science	Physical Science
Men	547.7(3.2)	513.4(2.0)	505.0(2.5)	534.8(1.5)
Women	535.2(2.4)	481.2(1.4)	482.6(1.8)	522.0(2.4)
M-W Average	541.4(2.0)	497.3(1.2)	493.8(1.5)	528.4(1.4)
M-W Difference	12.5(4.0)	32.2(2.4)	22.4(3.1)	12.8(2.8)

One-way analysis of variance: $F = 172$; $ETA^2 = .050$

Table 3d

Mean GRE-quantitative Scores by Sex Within Field of Study with Across-Sex Average Scores and Male-Female Differences Within Field of Study (Standard Errors in Parentheses)

	Humanities	Social Science	Biological Science	Physical Science
Men	575.2(3.2)	557.4(2.1)	598.3(2.5)	686.5(1.2)
Women	517.4(2.4)	499.1(1.5)	535.0(2.0)	645.2(2.2)
M-W Average	546.3(2)	528.2(1.3)	566.6(1.6)	665.8(1.3)
M-W Difference	57.8(4)	58.4(2.6)	63.3(3.2)	41.3(2.5)

One-way analysis of variance: F = 1,459; ETA^2 = .308

Table 3e

Mean GRE-analytical Scores by Sex Within Field of Study with Across-Sex Average Scores and Male-Female Differences Within Field of Study (Standard Errors in Parentheses)

	Humanities	Social Science	Biological Science	Physical Science
Men	579.8(3.3)	561.5(2.2)	579.3(2.8)	633.5(1.6)
Women	568.2(2.4)	542.3(1.6)	562.3(2.0)	630.1(2.6)
M-W Average	574.0(2.0)	551.9(1.4)	570.8(1.7)	631.8(1.5)
M-W Difference	11.6(4.1)	19.2(2.7)	17.0(3.4)	3.4(3.1)

One-way analysis of variance: F = 317; ETA^2 = .088

Table 4

Coefficients of Within-Group Linear Prediction of GRE-verbal Scores from SAT-verbal Scores

Field of Study	Sex	Intercept	SAT-V Slope*	S.E. of Estimate	R^2	SAT-M Contrib. to R^2**
Humanities	Men	27.0	.950(.015)	57.3	.754	.004
	Women	21.3	.932(.012)	55.2	.751	.005
Social Science	Men	53.0	.902(.010)	57.6	.734	.006
	Women	59.1	.856(.007)	53.4	.742	.008
Biological Science	Men	87.7	.833(.014)	54.1	.692	.009
	Women	77.3	.818(.010)	53.4	.694	.010
Physical Science	Men	51.6	.889(.008)	55.0	.713	.003
	Women	42.0	.886(.013)	53.0	.735	.008

*Standard Error of slope in parentheses.

**Increase in R^2 by adding SAT-mathematical score to the regression.

Table 5

Coefficients of Within-Group Linear Prediction of GRE-quantitative Scores from SAT-mathematical Scores

Field of Study	Sex	Intercept	SAT-M Slope*	S.E. of Estimate	R^2	SAT-V Contrib. to R^2**
Humanities	Men	42.8	.927(.015)	58.7	.734	.001
	Women	15.7	.943(.014)	61.4	.695	.003
Social Science	Men	53.3	.932(.011)	62.6	.716	.002
	Women	35.3	.928(.009)	62.7	.682	.004
Biological Science	Men	119.4	.855(.015)	58.1	.663	.004
	Women	54.2	.923(.012)	61.1	.680	.012
Physical Science	Men	216.3	.734(.008)	50.5	.630	.004
	Women	151.5	.815(.014)	53.1	.661	.006

*Standard Error of slope in parentheses.

**Increase in R^2 by adding SAT-verbal score to the regression.

Table 6

Predicted Values of GRE-verbal Scores for SAT-verbal Scores at the 10th and 90th Percentiles*, by Subgroup (Standard Errors in Parentheses)

10th Percentile: SAT-verbal = 380

	Humanities	Social Science	Biological Science	Physical Science	Range
Men	388(3.0)	396(1.7)	403(2.2)	389(1.6)	15(3.7)
Women	376(2.3)	384(1.1)	388(1.5)	379(2.4)	12(2.8)
M-W Average	382(1.9)	390(1.0)	396(1.3)	384(1.4)	14(2.3)
M-W Difference	12(3.8)	12(2.0)	15(2.7)	10(2.9)	5(4.0)

90th Percentile: SAT-verbal = 650

	Humanities	Social Science	Biological Science	Physical Science	Range
Men	645(2.2)	639(1.7)	628(2.5)	629(1.2)	17(3.3)
Women	627(1.7)	616(1.3)	609(1.8)	618(1.9)	18(2.5)
M-W Average	636(1.4)	628(1.1)	619(1.5)	624(1.1)	17(2.1)
M-W Difference	18(2.8)	23(2.1)	19(3.1)	11(2.3)	12(3.1)

*Percentiles are based on the students in the total study sample.

Table 7

Predicted Values of GRE-quantitative Scores for SAT-mathematical Scores at the 10th and 90th Percentiles*, by Subgroup (Standard Errors in Parentheses)

10th Percentile: SAT-mathematical = 400

	Humanities	Social Science	Biological Science	Physical Science	Range
Men	413(3.2)	426(1.9)	461(2.9)	510(2.1)	97(3.8)
Women	393(2.2)	407(1.2)	423(1.8)	477(3.1)	84(3.8)
M-W Average	403(1.9)	417(1.1)	442(1.7)	494(1.9)	91(2.7)
M-W Difference	20(3.9)	19(2.3)	38(3.4)	33(3.7)	19(4.1)

90th Percentile: SAT-mathematical = 700

	Humanities	Social Science	Biological Science	Physical Science	Range
Men	691(2.5)	706(2.0)	718(2.6)	730(0.9)	39(2.7)
Women	675(2.6)	685(1.9)	700(2.4)	722(1.8)	47(3.2)
M-W Average	683(1.8)	696(1.4)	709(1.8)	726(1.0)	43(2.1)
M-W Difference	16(3.6)	21(2.8)	18(3.5)	8(2.0)	13(3.4)

*Percentiles are based on the students in the total study sample.

Table 8

Means and Standard Deviations of Observed GRE-verbal Scores, by Subgroup, for Given Ranges of SAT-verbal Score

		SAT-verbal Score											
		351-450			451-550			551-650			651-750		
		N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Humanities	Men	211	422.5	55.39	379	505.8	66.80	426	597.6	70.11	208	702.9	49.51
	Women	289	416.5	54.67	657	493.7	60.42	799	577.2	66.30	329	674.6	62.55
Social Science	Men	757	428.6	57.24	1032	499.0	67.61	706	603.7	68.35	340	680.0	60.59
	Women	1469	411.0	50.36	1998	481.2	56.01	1070	567.9	63.60	399	665.5	56.86
Biological Science	Men	352	421.4	55.70	606	499.3	52.96	405	578.7	64.25	91	683.0	55.98
	Women	793	417.3	55.63	1169	482.0	55.26	596	572.1	60.96	159	659.0	47.74
Physical Science	Men	726	426.0	55.19	1597	501.5	59.13	1519	578.0	60.29	628	671.1	56.64
	Women	308	419.0	59.12	636	484.8	50.33	557	574.1	57.14	205	673.3	52.83

Table 9

Means and Standard Deviations of Observed GRE-quantitative Scores, by Subgroup, for Given Ranges of SAT-mathematical Score

SAT-mathematical Score

		351-450			451-550			551-650			651-750		
		N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Humanities	Men	154	415.1	74.11	377	515.7	69.44	453	602.1	64.49	260	683.7	54.70
	Women	353	405.9	81.90	801	492.9	73.08	655	584.0	63.56	244	668.5	55.54
Social Science	Men	576	436.8	74.42	953	521.0	79.14	901	620.1	67.21	418	709.6	61.50
	Women	1366	417.5	68.95	2122	501.1	64.71	1262	587.3	65.80	325	675.6	51.93
Biological Science	Men	162	470.7	80.35	554	558.4	67.95	544	644.4	53.71	226	711.6	45.91
	Women	602	435.6	72.85	1128	527.5	69.0	848	605.0	64.28	219	677.1	46.81
Physical Science	Men	158	538.2	86.52	551	583.8	64.87	1631	668.8	57.10	1823	733.6	43.68
	Women	107	497.8	77.04	361	583.2	69.42	735	648.9	62.19	514	707.9	49.19

Table 10

Coefficients of Within-Group Linear Prediction
of GRE-analytical Scores from
SAT-verbal and SAT-mathematical Scores

Field of Study	Sex	Intercept	SAT-V Slope*	SAT-M Slope*	S.E. of Estimate	R^2	Contribution to R^2 by: SAT-V	SAT-M
Humanities	Men	58.5	.339(.026)	.584(.026)	77.6	.566	.056	.165
	Women	88.4	.321(.020)	.569(.021)	76.4	.524	.057	.164
Social Science	Men	67.5	.306(.018)	.624(.017)	77.3	.582	.042	.177
	Women	75.5	.361(.013)	.578(.014)	75.2	.573	.063	.140
Biological Science	Men	86.6	.314(.025)	.597(.025)	75.5	.527	.048	.168
	Women	88.0	.346(.018)	.581(.018)	72.3	.567	.055	.148
Physical Science	Men	61.5	.316(.014)	.625(.015)	73.3	.536	.051	.170
	Women	87.5	.336(.022)	.595(.024)	73.6	.536	.061	.160

* Standard Error of slope in parentheses

Table 11

Predicted Values of GRE-analytical Scores
for SAT-verbal and SAT-mathematical Scores at the 10th and 90th Percentiles*,
by Subgroup

(Standard Errors in Parentheses)

10th Percentiles: SAT-verbal = 380; SAT-mathematical = 400

	Humanities	Social Science	Biological Science	Physical Science	Range
Men	421(4.4)	434(2.4)	445(3.8)	432(3.1)	24(5.8)
Women	438(3.3)	444(1.5)	452(2.2)	453(4.4)	15(5.5)
M-W Average	430(2.8)	439(1.4)	449(2.2)	443(2.7)	19(3.6)
M-W Difference	−17(5.5)	−10(2.8)	−7(4.4)	−21(5.4)	14(7.0)

90th Percentiles: SAT-verbal = 650; SAT-mathematical = 700

	Humanities	Social Science	Biological Science	Physical Science	Range
Men	688(3.4)	704(2.6)	709(3.7)	704(1.6)	21(5.0)
Women	695(3.3)	715(2.3)	720(2.9)	722(2.7)	27(4.3)
M-W Average	692(2.4)	710(1.7)	715(2.4)	713(1.6)	23(3.4)
M-W Difference	−7(4.7)	−11(3.5)	−11(4.7)	−18(3.1)	11(5.6)

*Percentiles are based on the students in the total study sample.

Table 12

Predicted Values of GRE-verbal Scores
for the "Verbal" Humanities and "Mathematical" Physical Science Fields of Study
for SAT-verbal Scores at the 10th and 90th Percentiles*

(Standard Errors in Parentheses)

10th Percentile: SAT-verbal = 380

	"Verbal" Humanities	"Mathematical" Physical Science	Difference
Men	418(7.5)	391(2.9)	27(8.0)
Women	389(4.9)	380(4.7)	9(6.8)
M-W Average	404(4.5)	386(2.8)	18(5.3)
M-W Difference	29(9.0)	11(5.5)	18(10.6)

90th Percentile: SAT-verbal = 650

	"Verbal" Humanities	"Mathematical" Physical Science	Difference
Men	656(4.5)	632(2.1)	24(5.0)
Women	636(3.2)	625(4.0)	11(5.1)
M-W Average	646(2.8)	629(2.3)	17(3.6)
M-W Difference	20(5.5)	7(4.5)	13(7.1)

*Percentiles are based on the students in the total study sample.

Table 13

Predicted Values of GRE-quantitative Scores
for the "Verbal" Humanities and "Mathematical" Physical Science Fields of Study
for SAT-mathematical Scores at the 10th and 90th Percentile*

(Standard Errors in Parentheses)

10th Percentile: SAT-mathematical = 400

	"Verbal" Humanities	"Mathematical" Physical Science	Difference
Men	426(7.8)	536(3.9)	-110(8.7)
Women	390(4.3)	492(5.9)	-102(7.3)
M-W Average	408(4.5)	514(3.5)	-106(5.7)
M-W Difference	36(8.9)	44(7.1)	-8(11.4)

90th Percentile: SAT-mathematical = 700

	"Verbal" Humanities	"Mathematical" Physical Science	Difference
Men	681(6.2)	734(1.5)	-53(6.4)
Women	673(5.3)	730(3.3)	-57(6.2)
M-W Average	677(4.1)	732(1.8)	-55(4.5)
M-W Difference	8(8.2)	4(3.6)	4(9.0)

*Percentiles are based on the students in the total study sample.

Table 14

Predicted Values of GRE-analytical Scores
for the "Verbal" Humanities and "Mathematical" Physical Science Fields of Study
for SAT-verbal and SAT-mathematical Scores at the 10th and 90th Percentiles*

(Standard Errors in Parentheses)

10th Percentiles: SAT-verbal = 380; SAT-mathematical = 400

	"Verbal" Humanities	"Mathematical" Physical Science	Difference
Men	448(10.2)	433(5.9)	15(11.8)
Women	432(6.7)	450(9.4)	-18(11.5)
M-W Average	440(6.1)	442(5.6)	-2(8.3)
M-W Difference	16(12.2)	-17(11.1)	33(16.5)

90th Percentiles: SAT-verbal = 650; SAT-mathematical = 700

	"Verbal" Humanities	"Mathematical" Physical Science	Difference
Men	683(7.3)	710(2.8)	-27(7.8)
Women	707(6.6)	729(5.7)	-22(8.7)
M-W Average	695(4.9)	720(3.2)	-25(5.9)
M-W Difference	-24(9.8)	-19(6.4)	-5(11.7)

*Percentiles are based on the students in the total study sample.

Table 15

Variance Component Analysis
for form K-3FGR3 of the GRE-quantitative Test*

Fixed-effect parameters
from multiple regression equation

Variables and additive adjustments		Estimate	Standard Error
Intercept		72.0	0.9
Major:	Humanities	0.0	0.0
	Social Sciences	−12.3	12.3
	Biological Sciences	20.0	15.0
	Physical Sciences	98.5	21.8
Sex:	Men	0.0	0.0
	Women	−17.9	1.4
SAT-mathematical score		0.87	0.02
SAT-M by Major:			
	Humanities	0.00	0.00
	Social Sciences	0.048	0.024
	Biological Sciences	0.027	0.029
	Physical Sciences	−0.042	0.040
SATMSQ = $(\text{SAT-M} - 500)^2/100$		2.33	1.29
SATMSQ by Major:			
	Humanities	0.00	0.00
	Social Science	−2.35	1.53
	Biological Science	−4.18	1.85
	Physical Science	−7.32	1.79

Variance Components

Source	Component	Variance	Sigma**
Student	σ^2_ε	3225.1	56.8
Institution	σ^2_I (Intercept)	112.2	10.6 (0.9)
	σ^2_S (SAT-M slope)	0.0012	0.0346(0.0109)

σ_{IS} = Covariance between intercept and SAT-M slope = −.3658 (.0584)

*Based on 7954 students from the 292 institutions with at least 10 students taking each form.

**Standard errors in parentheses.

Table 16

Descriptive Statistics from the Mantel-Haenszel Analyses,
Referencing Each Subgroup Against
Men in the Humanities

GRE-verbal; 76 items

Focal Group		Selected Order Statistics of MH D-DIF: Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Mean D-DIF*	Mean-Squared Errors of D-DIF: Within Items	Between Items
Humanities	Women	-2.56	-.41	-.13	.34	2.12	-.13(.01)	.14	.56
Social Science	Men	-1.03	-.29	.03	.37	1.55	.06(.01)	.12	.26
	Women	-2.49	-.51	-.09	.30	1.52	-.10(.01)	.11	.41
Biological Science	Men	-1.61	-.55	-.17	.35	2.63	.02(.01)	.15	.65
	Women	-2.37	-.74	-.29	.22	2.08	-.22(.01)	.12	.73
Physical Science	Men	-2.29	-.74	-.29	.24	1.54	-.29(.01)	.13	.53
	Women	-1.99	-.92	-.38	.06	1.13	-.43(.01)	.15	.55
Social Science	M-W Average	-1.41	-.31	-.03	.22	1.26	-.02(.01)	.06	.20
	M-W Difference	-1.54	-.33	.02	.59	2.17	.16(.02)	.23	.52
Biological Science	M-W Average	-1.44	-.55	-.22	.15	2.36	-.10(.01)	.07	.56
	M-W Difference	-1.49	-.15	.14	.53	3.17	.23(.02)	.29	.53
Physical Science	M-W Average	-1.84	-.75	-.32	-.04	1.33	-.36(.01)	.07	.42
	M-W Difference	-2.13	-.22	.04	.55	2.39	.15(.02)	.27	.49

*Standard error in parentheses

Table 17

Descriptive Statistics from the Mantel-Haenszel Analyses,
Referencing Each Subgroup Against
Men in the Humanities

GRE-quantitative; 60 items

		Selected Order Statistics of MH D-DIF						Mean-Squared Errors of D-DIF	
Focal Group		Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Mean D-DIF*	Within Items	Between Items
Humanities	Women	-2.09	-.69	-.26	-.06	.80	-.37(.01)	.12	.32
Social Science	Men	-.44	.13	.36	.71	1.36	.40(.01)	.11	.17
	Women	-1.13	-.43	-.04	.28	1.33	-.02(.01)	.09	.31
Biological Science	Men	-.27	.50	.88	1.50	4.10	.99(.01)	.16	.64
	Women	-1.08	-.22	.21	.66	1.91	.26(.01)	.11	.47
Physical Science	Men	.05	.76	1.58	2.17	4.55	1.55(.02)	.15	1.11
	Women	-.66	.28	1.20	1.96	5.06	1.17(.02)	.18	1.51
Social Science	M-W Average	-.66	-.09	.15	.42	1.34	.19(.01)	.05	.18
	M-W Difference	-.59	.04	.45	.70	2.18	.42(.02)	.20	.29
Biological Science	M-W Average	-.53	.13	.57	1.13	2.70	.63(.01)	.07	.45
	M-W Difference	-.60	.31	.66	1.07	2.81	.72(.02)	.29	.40
Physical Science	M-W Average	-.23	.52	1.13	2.03	4.33	1.36(.01)	.08	1.21
	M-W Difference	-1.60	.06	.39	.81	1.05	.41(.02)	.32	.41

*Standard error in parentheses

Table 18

Descriptive Statistics from the Mantel-Haenszel Analyses, Referencing Each Subgroup Against Men in the Humanities

GRE-analytical; 50 items

Focal Group		Selected Order Statistics of MH D-DIF: Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Mean D-DIF*	Mean-Squared Errors of D-DIF: Within Items	Between Items
Humanities	Women	-1.42	-.21	-.23	.59	1.24	.17(.01)	.12	.31
Social Science	Men	-.56	-.03	.27	.40	.82	.19(.01)	.10	.12
	Women	-1.00	-.01	.37	.80	1.92	.42(.01)	.09	.29
Biological Science	Men	-.43	.18	.57	.77	2.35	.54(.01)	.14	.25
	Women	-.38	.07	.47	.93	2.55	.51(.01)	.11	.29
Physical Science	Men	-.56	.05	.28	.60	1.21	.31(.01)	.12	.15
	Women	-1.03	.28	.64	1.01	2.17	.64(.01)	.15	.38
Social Science	M-W Average	-.34	.05	.30	.60	1.37	.31(.01)	.08	.14
	M-W Difference	-1.11	-.63	-.26	.19	1.39	-.22(.02)	31	.25
Biological Science	M-W Average	-.27	.26	.53	.76	2.08	.52(.01)	.06	.18
	M-W Difference	-.94	-.39	-.08	.31	2.52	.02(.02)	.24	.35
Physical Science	M-W Average	-.43	.20	.44	.78	1.69	.48(.01)	.07	.18
	M-W Difference	-1.55	-.69	-.41	-.02	2.01	-.33(.02)	.28	.32

*Standard error in parentheses

Table 19

Counts of Items by Type of Result from the Mantel-Haenszel Analyses, Referencing Each Subgroup Against Men in the Humanities

GRE-verbal; 76 items

Focal Group		Favoring Reference Group			Favoring Focal Group	
		$MH^* \le -2$	$-2 < MH^* \le -1$	$\lvert MH^* \rvert < 1$	$1 \le MH^* < 2$	$MH^* \ge 2$
Humanities	Women	3	4	68	0	1
Social Science	Men	0	1	71	4	0
	Women	1	6	66	3	0
Biological Science	Men	0	5	60	10	1
	Women	1	11	57	6	1
Physical Science	Men	2	10	61	3	0
	Women	0	18	57	1	0

*MH = MH D-DIF

Table 20

Counts of Items by Type of Result
from the Mantel-Haenszel Analyses,
Referencing Each Subgroup Against
Men in the Humanities

GRE-quantitative; 60 items

Focal Group		Favoring Reference Group			Favoring Focal Group	
		$MH^* \leq -2$	$-2 < MH^* \leq -1$	$\lvert MH^* \rvert < 1$	$1 \leq MH^* < 2$	$MH^* \geq 2$
Humanities	Women	1	8	51	0	0
Social Science	Men	0	0	55	5	0
	Women	0	3	54	3	0
Biological Science	Men	0	0	34	22	4
	Women	0	1	51	8	0
Physical Science	Men	0	0	24	18	18
	Women	0	0	28	18	14

*MH = MH D-DIF

Table 21

Counts of Items by Type of Result
from the Mantel-Haenszel Analyses,
Referencing Each Subgroup Against
Men in the Humanities

GRE-analytical; 50 items

Focal Group		Favoring Reference Group			Favoring Focal Group	
		$MH^* \le -2$	$-2 < MH^* \le -1$	$\lvert MH^* \rvert < 1$	$1 \le MH^* < 2$	$MH^* \ge 2$
Humanities	Women	0	2	44	4	0
Social Science	Men	0	0	50	0	0
	Women	0	1	44	5	0
Biological Science	Men	0	0	45	4	1
	Women	0	0	40	9	1
Physical Science	Men	0	0	48	2	0
	Women	0	1	37	12	0

*MH = MH D-DIF

Table 22

Numbers of Items Whose Overall Statistic
Exceeds Selected Percentiles
in a Chi-Square Distribution (7 df)

	Verbal	Quantitative	Analytical
Number ≥ 90th percentile	9	22	1
Number ≥ 95th percentile	4	17	1
Number ≥ 99th percentile	3	11	1
Total Nos. of Items	76	60	50

FIGURE 1

Plot of Bivariate Means
for the Subgroups of the Study Sample:
SAT-verbal vs GRE-verbal

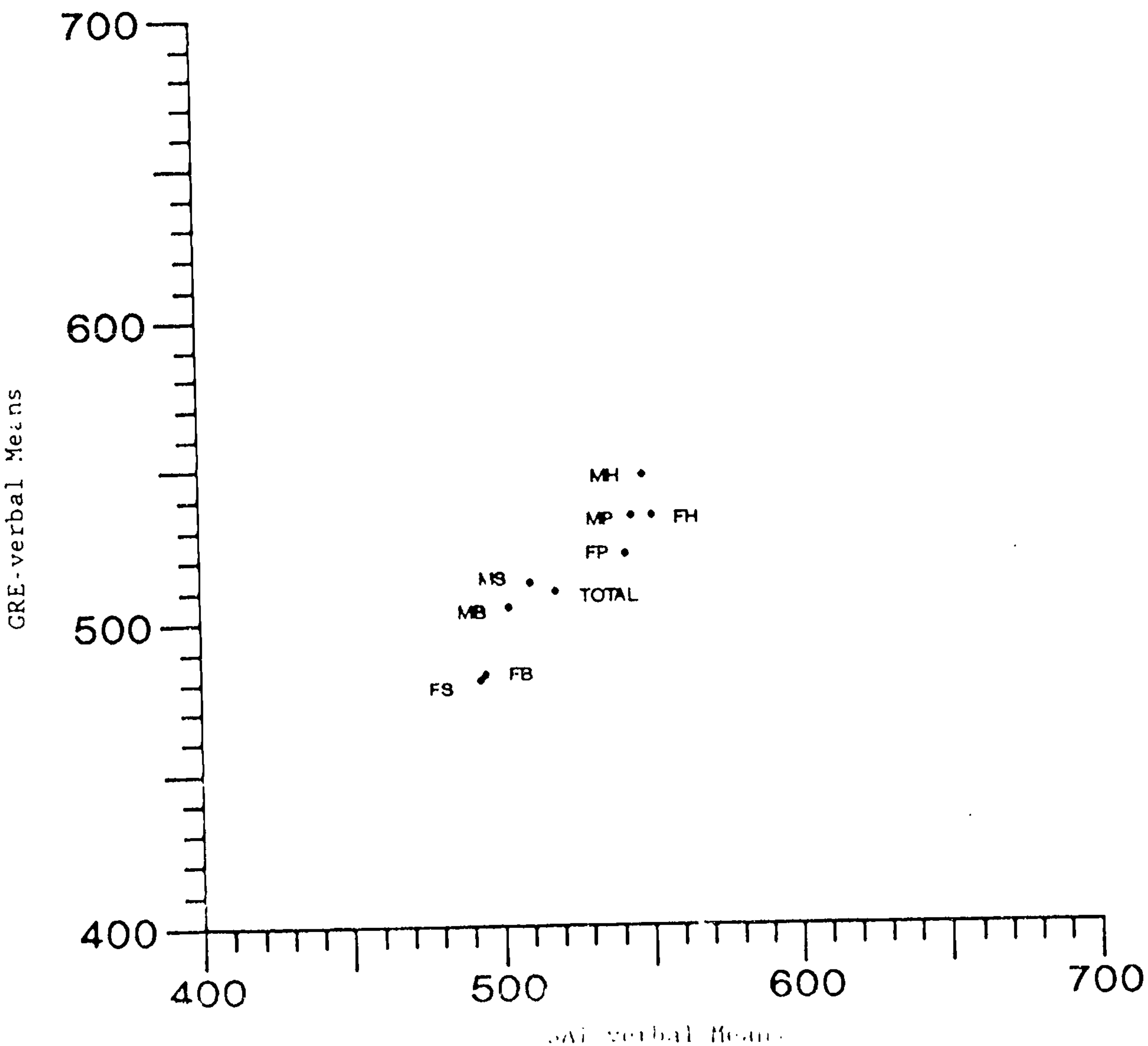


FIGURE 2

Plot of Bivariate Means
For the Subgroups of the Study Sample:
SAT-mathematical vs GRE-quantitative

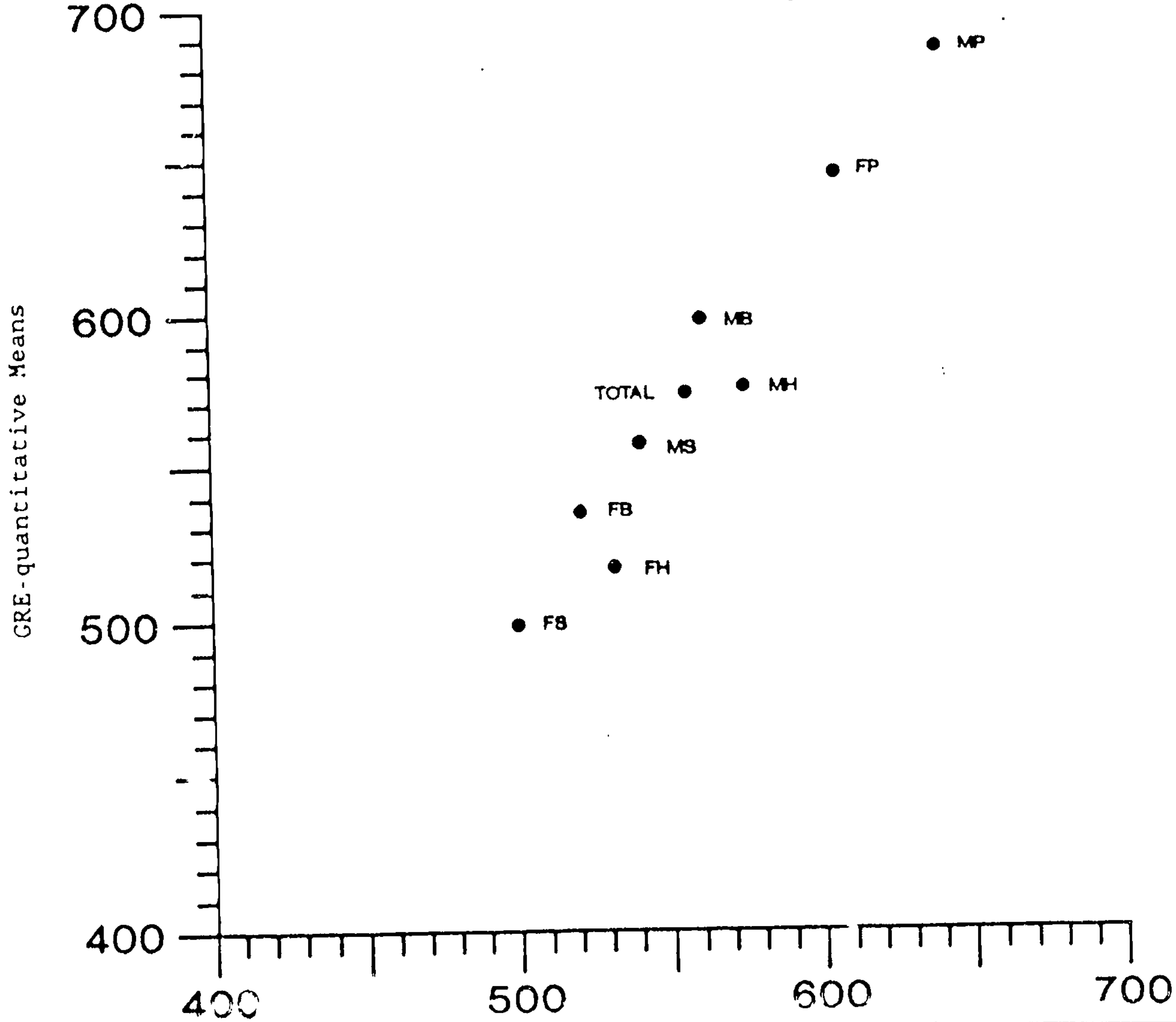


FIGURE 3

Regression Lines, by Subgroup,
for Predicting GRE-verbal Scores
from SAT-verbal Scores

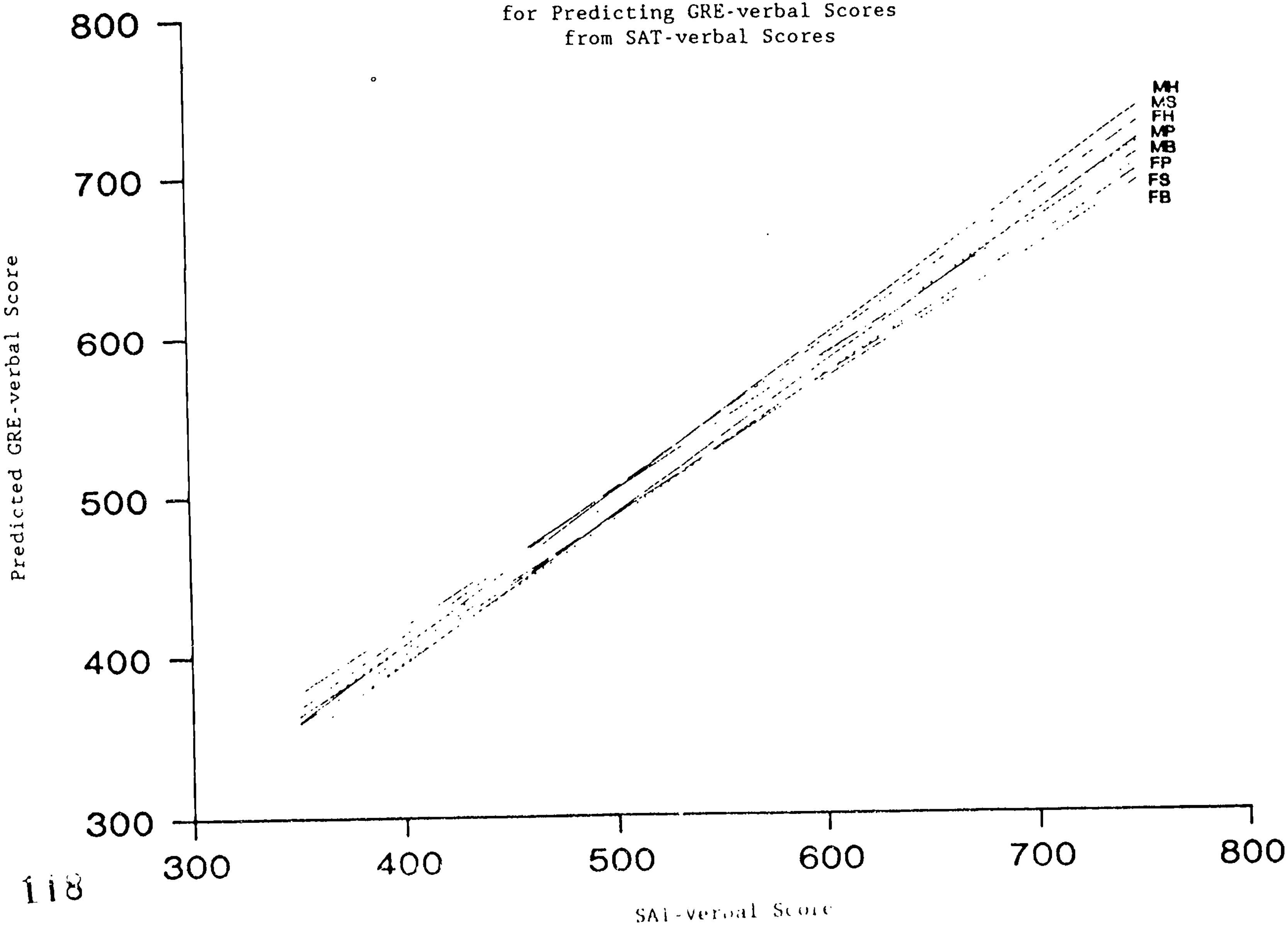


FIGURE 4

Regression Lines, by Subgroup,
for Predicting GRE-quantitative Scores
from SAT-mathematical Scores

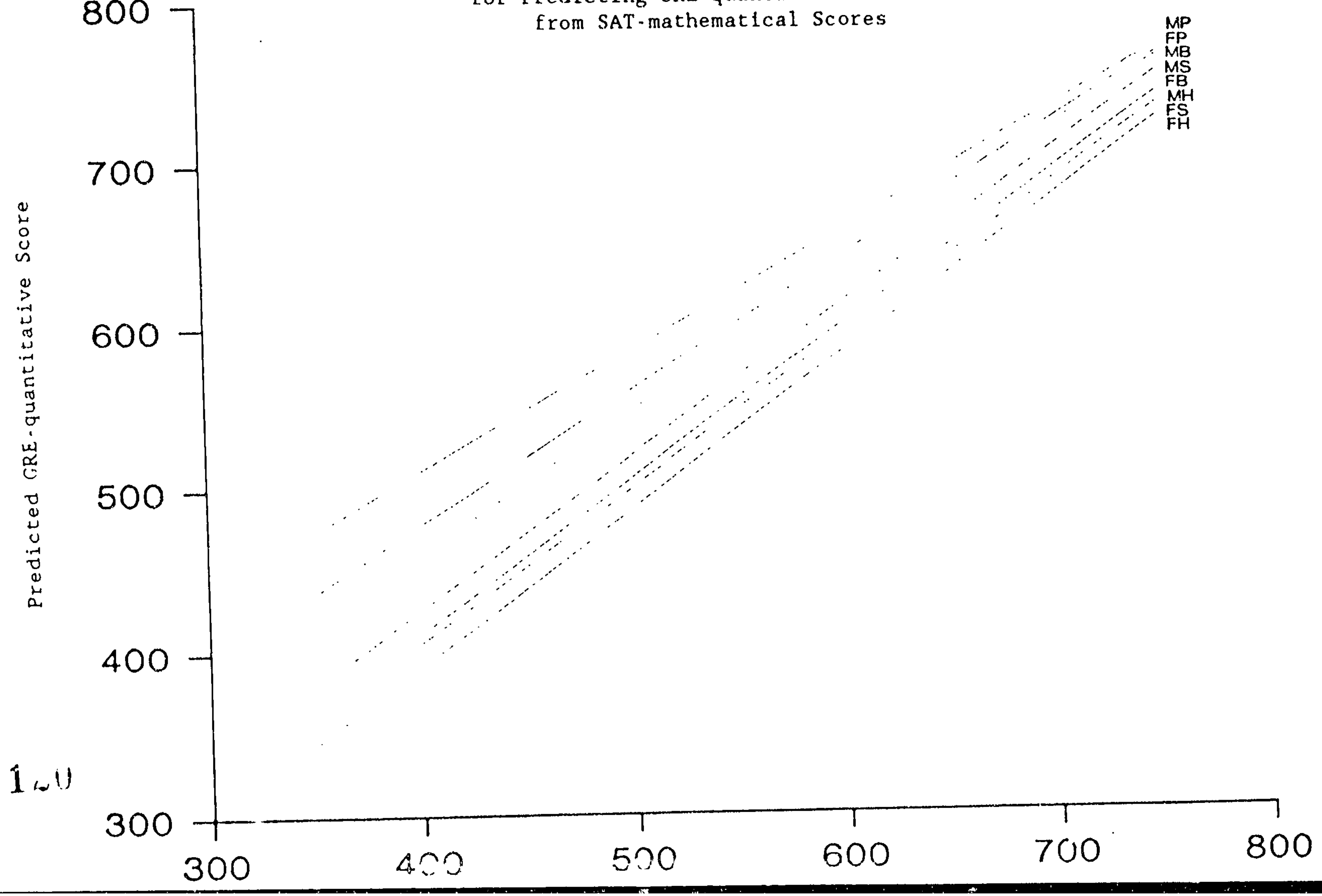


FIGURE 5

Regression Lines for Predicting GRE-verbal Scores
from SAT-verbal Scores,
as Deviations from the Average Regression Line

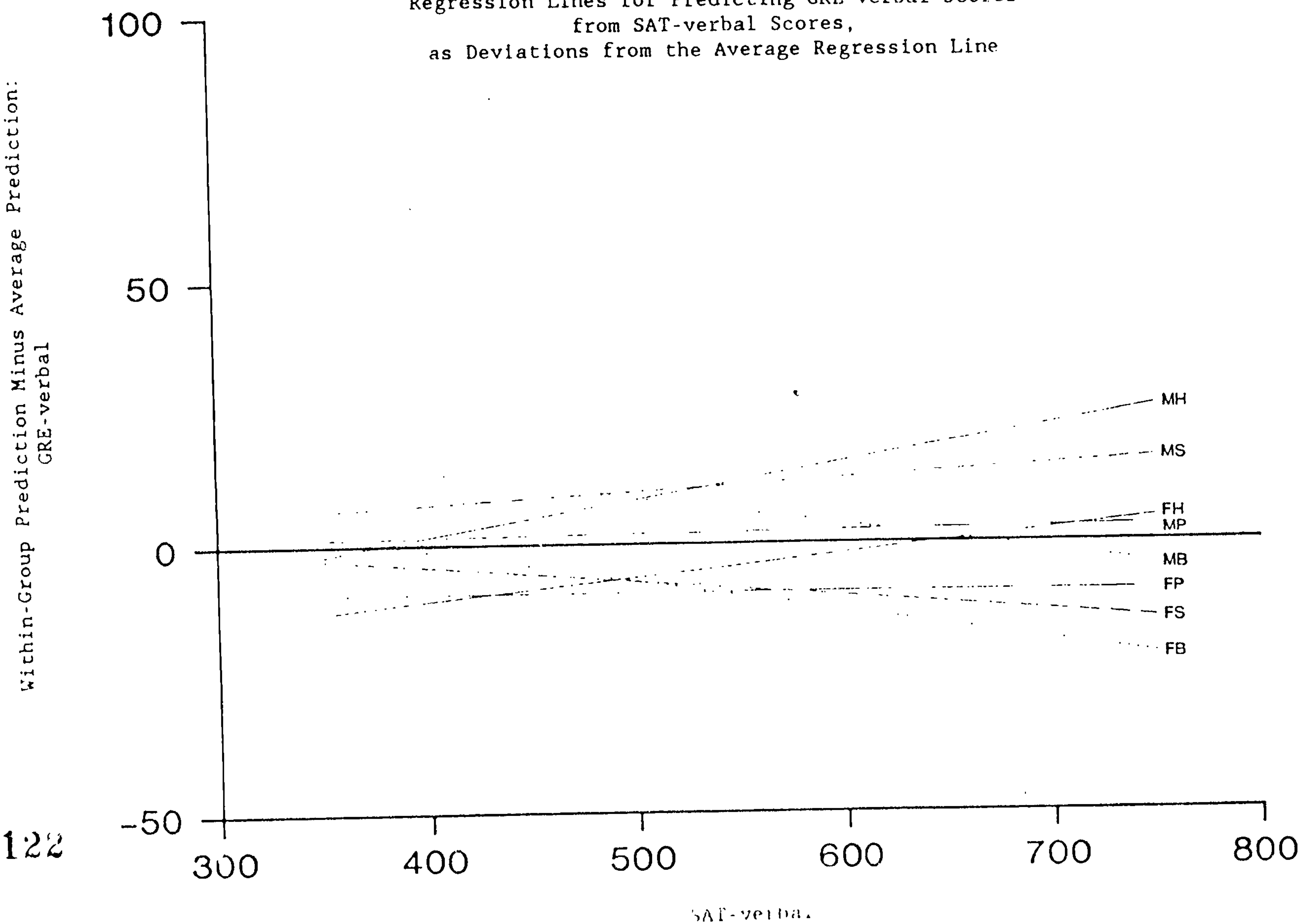


FIGURE 6

Regression Lines for Predicting GRE-quantitative Scores
from SAT-mathematical Scores
as Deviations from the Average Regression Line

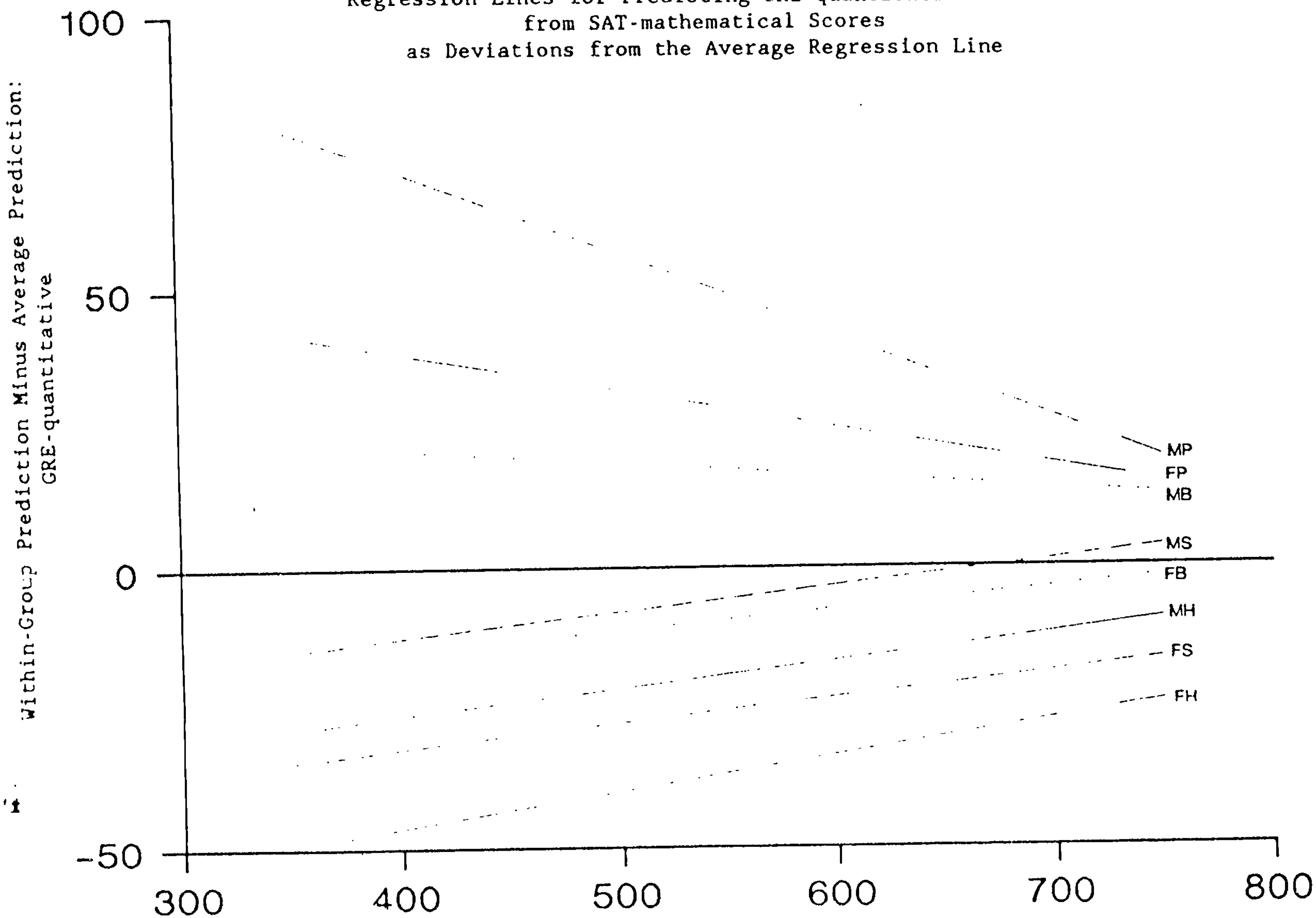


FIGURE 7

Mean Residuals (Actual minus Predicted)
for GRE-verbal Scores

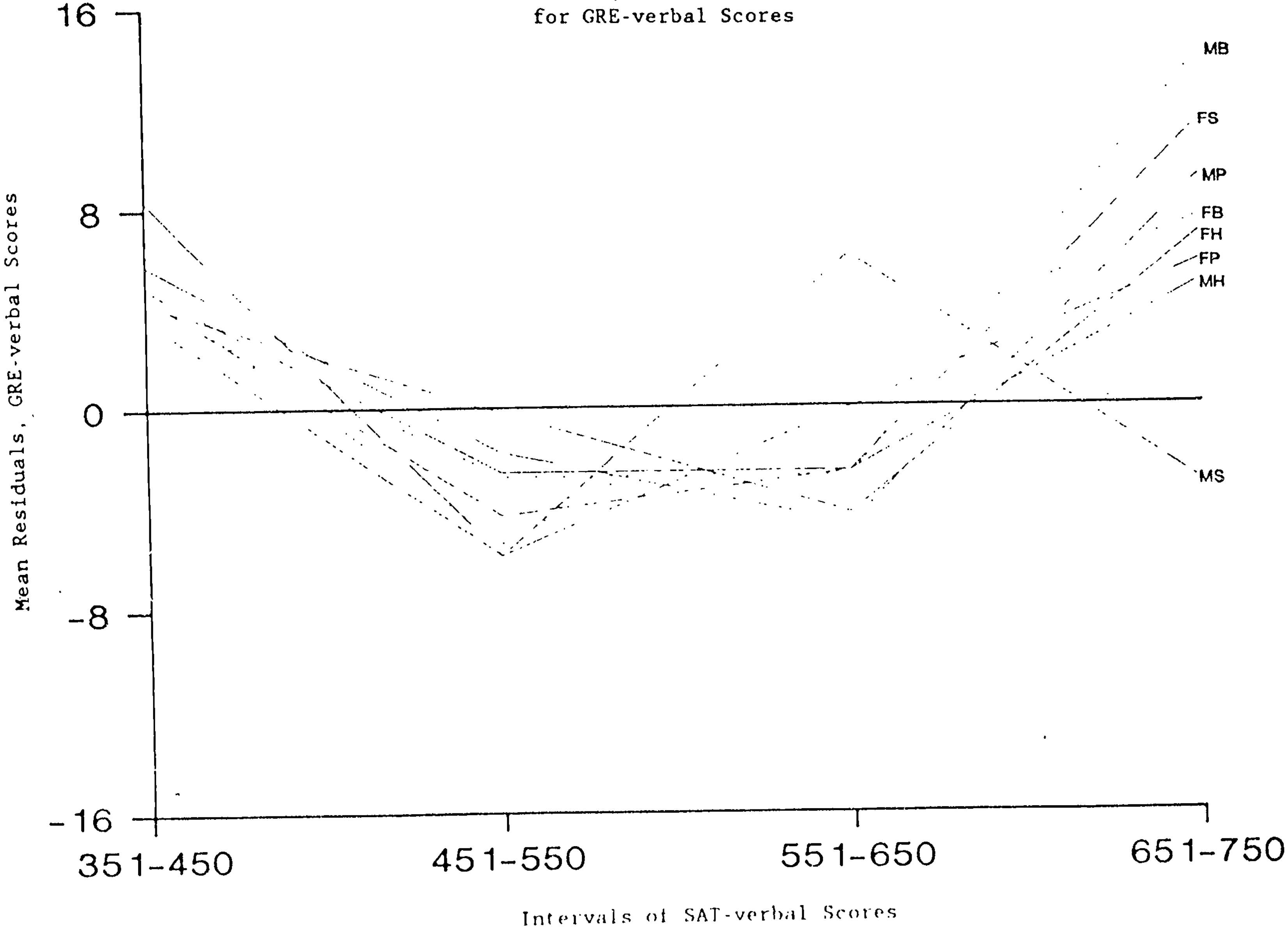


FIGURE 8

Mean Residuals (Actual minus Predicted)
for GRE-quantitative Scores

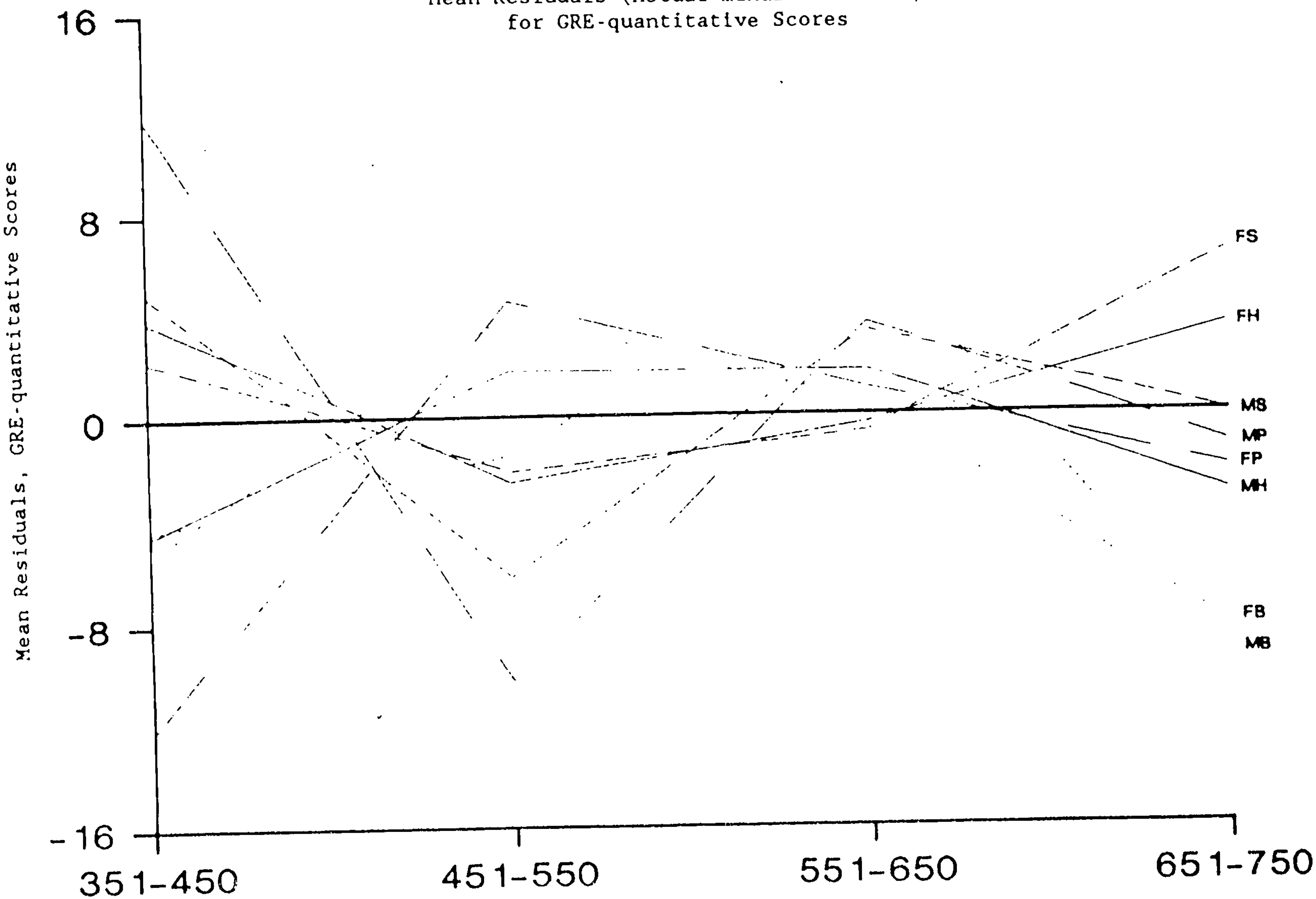


FIGURE 9

Regression Lines, by Subgroup,
for Predicting GRE-analytical Scores
from SAT-verbal and SAT-mathematical Scores

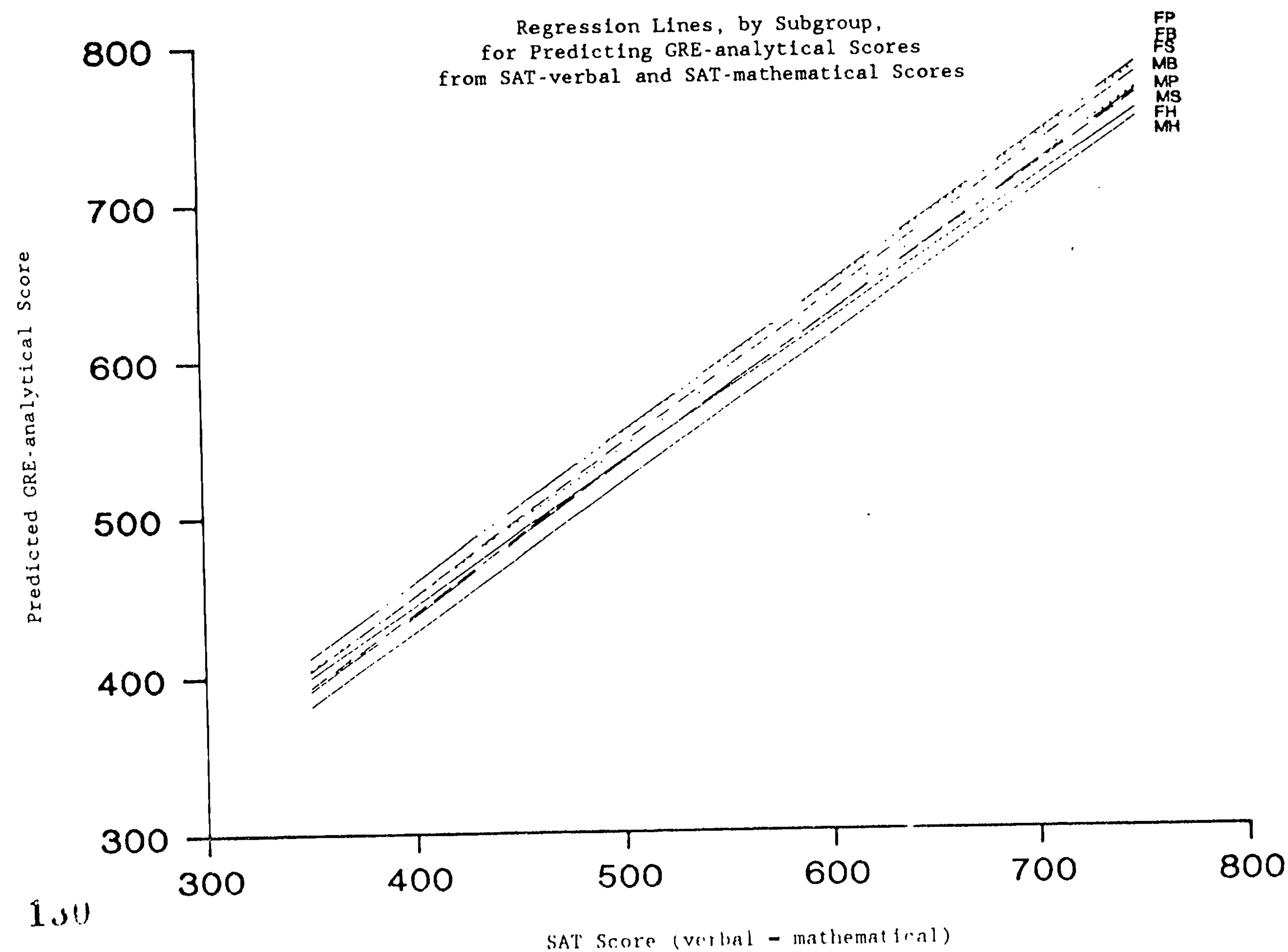
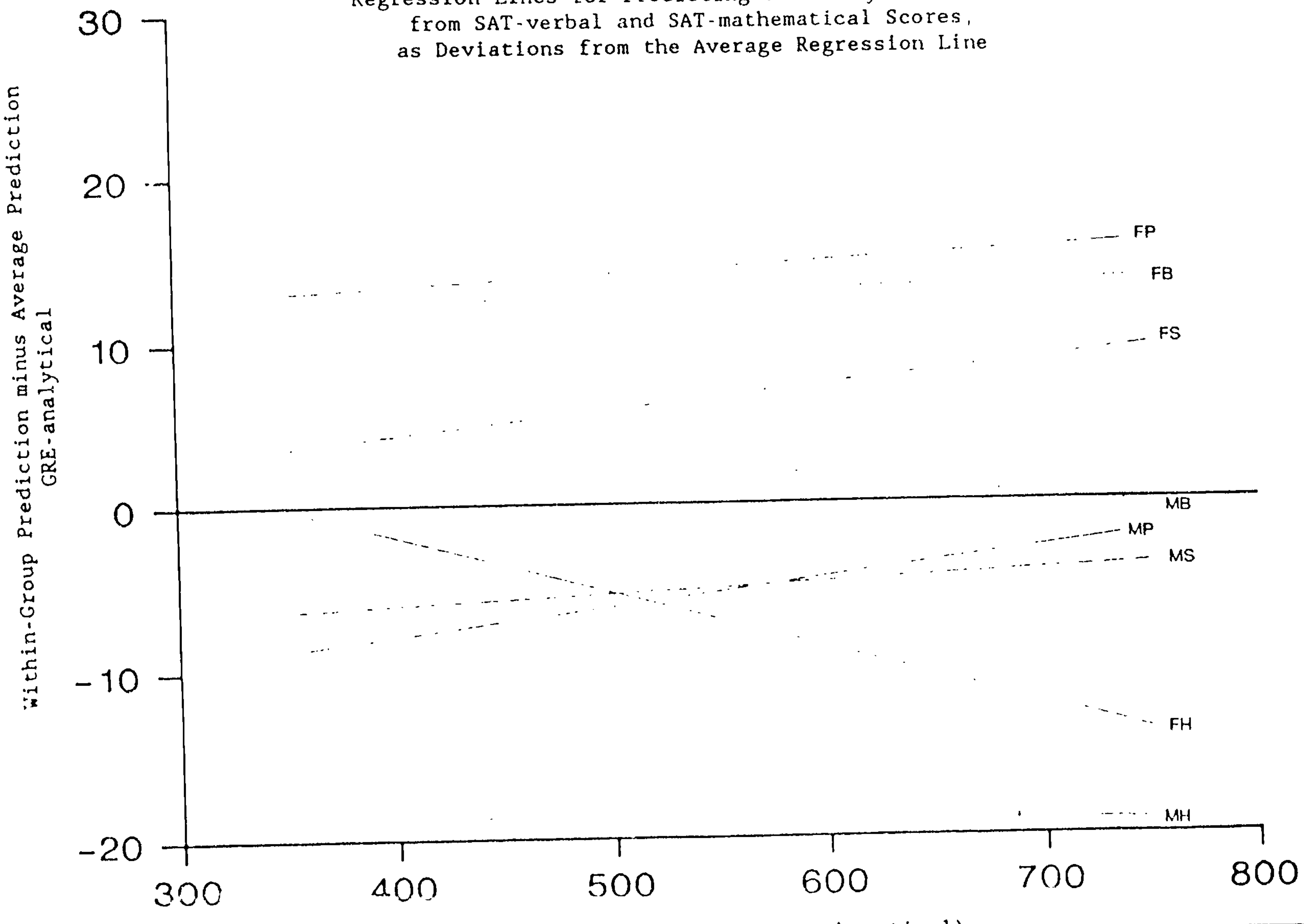


FIGURE 10

Regression Lines for Predicting GRE-analytical Scores
from SAT-verbal and SAT-mathematical Scores,
as Deviations from the Average Regression Line



APPENDIX

Major Field Code List

APPENDIX

Major Field Code List*

HUMANITIES
- Archaeology
- Architecture
- Art History
- Classical Languages
- Comparative Literature
- Dramatic Arts
- English
- Far Eastern Languages and Literature
- Fine Arts, Art, Design
- French
- German
- Linguistics
- Music
- Near Eastern Languages and Literature
- Philosophy
- Religious Studies or Religion
- Russian/Slavic Studies
- Spanish
- Speech
- Other Foreign Languages
- Other Humanities

SOCIAL SCIENCES
- American Studies
- Anthropology
- Business and Commerce
- Communications
- Economics
- Education (including M.A. in Teaching)
- Educational Administration
- Geography
- Government
- Guidance and Counseling
- History
- Industrial Relations and Personnel
- International Relations
- Journalism
- Law
- Library Science
- Physical Education
- Planning (City, Community, Urban, Regional)
- Political Science
- Psychology, Clinical
- Psychology, Educational
- Psychology, Experimental/Developmental
- Psychology, Other
- Psychology, Social
- Public Administration
- Social Work
- Sociology
- Other Social Sciences

BIOLOGICAL SCIENCES
- Agriculture
- Anatomy
- Audiology
- Bacteriology
- Biology
- Biomedical Sciences
- Biophysics
- Botany
- Dentistry
- Entomology
- Environmental Science/Ecology
- Forestry
- Genetics
- Home Economics
- Hospital and Health Services Administration
- Medicine
- Microbiology
- Molecular & Cellular Biology
- Nursing
- Nutrition
- Occupational Therapy
- Pathology
- Pharmacology
- Pharmacy
- Physical Therapy
- Physiology
- Public Health
- Speech-Language Pathology
- Veterinary Medicine
- Zoology
- Other Biological Sciences

PHYSICAL SCIENCES
- Applied Mathematics
- Astronomy
- Chemistry
- Computer Sciences
- Engineering, Aeronautical
- Engineering, Chemical
- Engineering, Civil
- Engineering, Electrical
- Engineering, Industrial
- Engineering, Mechanical
- Engineering, Other
- Geology
- Mathematics
- Metallurgy
- Oceanography
- Physics
- Statistics
- Other Physical Sciences

*Taken from the GRE 1984-85 Information Bulletin; p. 82