THE EFFECTIVENESS OF COACHING FOR THE SAT:

REVIEW AND REANALYSIS OF RESEARCH FROM THE FIFTIES TO THE FTC

by Samuel Messick

in collaboration with

Donald L. Alderman, William Angoff, Ann Jungeblut, Donald E. Powers, Donald Rock, Donald B. Rubin
Educational Testing Service • Princeton, NJ

AND

T.W.F. Stroud
Queens University • Kingston, Ontario
Contents

Preface ................................................................. v

I. Overview ......................................................... 1

II. Before the FTC Study:
The Context of Prior Findings on Coaching for the SAT ... 7

III. Beyond the FTC Study:
Critique of Assumptions and Inferences ...................... 31

IV. Estimation of Combined Coaching/Self-Selection
Effects in the FTC Study: Detailed Summary and
Elaboration of the Stroud Reanalysis ........................ 43

V. Disentangling Coaching Effects and Differential Growth:
Detailed Summary of the Rock Reanalysis .................... 53

VI. Implications for Educational and
Testing Policy and Practice ................................... 61

VII. References ...................................................... 69

VIII. Appendices

1. Critical Notes on the FTC Coaching Study .............. 75

2. Stroud Report ................................................ 97

3. Rock Report .................................................. 123
This report examines evidence and arguments about the effectiveness of coaching for the SAT. It views the issue as being much more complicated than the simplistic question of whether coaching works or not. Coaching in and of itself is not automatically to be either rejected or encouraged; it has to be analyzed and evaluated—it matters what materials and practices are involved, at what cost in student time and resources, and with what effect on student skills and attitudes as well as test scores.

The SAT measures developed abilities of verbal and mathematical reasoning and comprehension that are acquired gradually over many years of experience and use in both school and nonschool settings. By virtue of this gradual development, these intellective skills are relatively difficult to improve markedly through brief courses of intervention in the final year or two of high school when the SAT is typically taken. Since these abilities are learned in manifold ways through both instruction and experience, one would expect high quality instruction over extended periods of time to improve them and hence to increase SAT scores. Indeed, score gain across the high school years is the typical pattern exhibited by students taking the SAT. Since coaching at its best is a form of teaching, the key questions are whether the coaching experience is of sufficient quality and sufficient duration to yield significant skill improvement as well as score improvement over and above the experiential growth that would have occurred regardless of the coaching program. If significant improvement requires relatively large amounts of student time devoted to coaching, then the problem becomes expanded to include questions of the difference between coaching and instruction and of the instrumental role of comprehension and reasoning skills in school learning as well as their status as explicit objectives of school learning.

Thus the issue is not just whether coaching works or not, but how much student time devoted to what kinds of coaching experiences yield what level of score improvements in comparison with the level of experiential growth occurring without those coaching experiences. Moreover, since students with different personal and background characteristics often exhibit different performance characteristics and probably even learn in different
ways, we should be alert to the possibility that coaching programs, like other forms of teaching, may have differential effects for different kinds of students.

The second printing of this report incorporates the changes previously noted on an errata sheet prepared after the first printing.

Many individuals have contributed to this effort. We would like to thank Thomas Donlon, Garlie Forehand, and Winton Manning for their careful review of the manuscript; Nathaniel Hartshorne for his editorial comments; Rex Jackson and Stephen Ivens for their helpful suggestions on the presentation of arguments and of summary data; Lloyd Bond, Robert Glaser, and Robert Linn for their advice on research and policy implications; and, Anthony Bryk for his comments on one of the applications of the growth model included in the first printing of this report. Special thanks go to John Tukey for his general review of the analysis and its ramifications and, in particular, for his gentle insistence that student contact time in some way or other holds the key to understanding coaching effects.

Samuel Messick
Princeton, New Jersey
This report presents a critique and reanalysis of the Federal Trade Commission’s (FTC) study of commercial coaching for the Scholastic Aptitude Test (SAT). The FTC study is one of the largest studies of coaching ever done and one of the few studies of commercial coaching extant, so it merits careful examination. But it is not the only coaching study ever done nor is it free from problems of design, so it should be examined in the context of prior findings.

The first part of the report summarizes the major results of earlier studies in a way that draws special attention to the strengths and limitations of the various study designs. One of the most important of these design features is random assignment of examinees to coaching treatment groups and noncoaching control groups, for only with random assignment can we consider treatment effects to be independent of prior status on any of a host of personal or background characteristics. With random assignment, there are no systematic differences between the experimental and control groups initially and if effective control conditions are maintained, the only systematic difference that will eventuate is that one group will have received coaching and the other will not. In the absence of randomization, as is the case in the FTC study, there is an inevitable equivocality in the interpretation of the results because some unmeasured personal characteristics might have influenced both the student’s decision to participate in a coaching program and that program’s apparent effectiveness. That is, the same personal factors that led a student to attend coaching school may be responsible, at least in part, for subsequent SAT performance that appears to be the result of the coaching program. A number of factors that might lead a person to seek coaching may also by themselves explain why such a person would subsequently perform better than expected on the outcome measure or posttest; for example, a student might not have scored as well on the PSAT or SAT as he or his parents expected in light of high school grades or, in contrast, he might be highly motivated to earn a high score to compensate for a prosaic high school performance. Thus, the effects of self-selection are confounded with effects of the coaching treatment in nonrandomized
studies and, consequently, self-selection factors afford plausible rival explanations for the results, or for part of the results, that might otherwise be identified as coaching effects. In such non-randomized designs, researchers usually attempt to control statistically for those potential self-selection factors that have been measured and to analyze the data in alternate ways to assess the sensitivity or robustness of the findings under various plausible assumptions, but there is no way to adjust statistically for self-selection factors that have not been assessed.

An historical appraisal of the effects of coaching on the SAT is complicated because most of the investigations prior to the FTC study were concerned with diverse special preparation programs typically offered by secondary schools. Furthermore, some of these studies used nonrandomized designs and therefore are hampered by the same problems of interpretation that affect the FTC study; some were also poorly controlled and involved small samples. On balance, the average effect associated with participation in a coaching or special preparation program according to those earlier studies that included some type of control group was less than 10 points for the SAT-Verbal score (on a scale running from 200 to 800 points) and less than 15 points or so for the Math score. For example, two studies were conducted using random assignment of students to coaching and control conditions, one dealing with SAT-Math and the other with SAT-Verbal: In a study by Evans and Pike (1973) an average effect of slightly over 16 points for SAT-M was obtained for special preparation involving seven 3-hour sessions with 21 additional hours of homework; Alderman and Powers (1979) estimated an overall special preparation effect across eight secondary schools of about 8 points for SAT-V. However, for particular groups of students and for particular coaching treatments, estimated effects of over 20 points were also reported in various studies. Although substantive content of the coaching programs was not systematically evaluated as part of these studies, the smaller effects appear to be associated with short-term cramming or drill-and-practice and the larger effects, found more often for Math than Verbal, with longer-term intensive programs involving skill development. In addition, results in at least two earlier studies suggested possible interactions for SAT-M as a function of years of math taken and sex. Relatively consistent with this historical context, the FTC study found negligible effects for students attending one commercial coaching school and average score increases of about 20 to 30 points for both SAT-V and -M for students attending another school. Where
the coaching program entailed 10 four-hour sessions plus homework.

Because random assignment could not be employed in the FTC study, this 20- to 30-point effect for students attending one school is actually an estimate of the joint effects of coaching and self-selection. This confounding of treatment and self-selection effects is one of the major issues addressed in the critique of the FTC study that follows in Section III. It is impossible to determine with confidence whether the effects reported by the FTC study may be attributable in whole or in part to self-selection not controlled under the study design rather than to any impact of the coaching program as such. No statistical reanalysis of the data, given the study design, can fully solve this problem.

Nevertheless, in an effort to contribute further to an understanding of this complex issue of coaching effectiveness, we commissioned two major reanalyses of FTC data. The first one followed an analysis-of-covariance type of model similar to the FTC approach, but included additional background scores as covariates. Furthermore, this first reanalysis introduced an important refinement in that estimated effects were obtained relative to the regression line for noncoached students rather than for pooled coached and noncoached students as in the FTC analysis. This reanalysis covered all three commercial schools in the FTC data set, including one for which no analyses were undertaken by the FTC because of its small sample size. The overall results were similar to the general FTC findings: inconsistent and negligible effects for students at two schools and for students at the third school combined coaching and self-selection effects of about 20 to 35 points for both Verbal and Math scores. In addition, for students attending the two largest schools, this study investigated interactions between the size of effects and the characteristics or backgrounds of the students. Two interactions, one involving race and the other self-reported parental income, were uncovered for students at one of the coaching schools, the one exhibiting negligible effects overall, but not at the other and only for SAT-V, not for SAT-M. The sample of black students was very small and somewhat atypical, but black students at that one school exhibited larger coaching/self-selection effects than nonblacks. Similarly, independent of the race interaction, students reporting low parental income exhibited somewhat larger coaching/self-selection effects than those reporting high parental income.

The second reanalysis employed a statistical model which takes account of differential rates of growth in SAT scores over
time, if they occur, for the coached and noncoached groups. However, this analysis does not adjust for self-selection factors or other differences between the groups on background variables unrelated to differential growth. Because this analysis required three test scores for each student, it was undertaken only at the largest coaching school, the one for which the FTC found a correlation between coaching school attendance and SAT performance. Differential growth in the abilities measured by the SAT appeared to operate in these data more for the Verbal score than for the Math score. The resulting estimates of the combined coaching/self-selection effect, taking differential growth into account, were about 17 points for Verbal and 30 points for Math.

After reviewing earlier studies and the FTC study, the present report then considers the implications of the findings for testing practice and policy. One key issue is the extent to which increased test scores attributable to coaching may represent stable long-term improvements in the verbal and mathematical reasoning skills measured by the SAT or instead reflect the overcoming of inadvertent sources of test difficulty unrelated to these reasoning skills, such as difficulty associated with test anxiety and unfamiliarity with different item formats and test-taking strategies. No coaching study, however, has yet directly addressed this basic question of whether obtained score increases reflect stable ability improvements or increased test wiseness.

The SAT measures verbal and mathematical reasoning abilities that develop over many years of experience and use in both nonschool and school settings, and it is difficult to improve them with short-term interventions at seventeen or eighteen years of age, when the SAT is typically taken. Nonetheless, these abilities are learned, and effective learning experiences that facilitate their further development and result in test score increases should be welcomed. Such score increases reflective of improved abilities do not invalidate the SAT—on the contrary, they contribute to the test's construct validity as a measure of verbal and mathematical reasoning abilities and presumably also to its predictive utility, since the improved abilities should serve the student well in criterion situations (such as school or college learning) that entail verbal and quantitative reasoning by analogy, induction, or deduction. Issues of equity of access to coaching programs that may develop these abilities are basically similar to issues of equity of access to effective school programs or to effective nonschool learning experiences. Thus, any score increases that may represent real improvements in the underlying abilities measured
would obviously have implications for instruction and educational practice. No study to date has systematically addressed the question of what kind of instruction may be most effective over what period, for what kinds of students or groups of students, in improving SAT performance. Data that would permit a detailed comparison and evaluation of the commercial coaching programs covered in the FTC report have not been made available.

On the other hand, any score increases that may result from the reduction of construct-irrelevant difficulty on the test have implications for testing practice. Such sources of difficulty should be eliminated to the extent possible through astute test-construction procedures. For this reason also, prospective examinees should be provided with test familiarization materials, sample tests that can be taken at home, and other aids to effective test taking, such as those prepared and distributed by the College Board. Issues of equity of access to coaching programs that help overcome such extraneous test difficulty become important to the extent that student differences in test-taking skills per se substantially influence test scores.
Before the FTC Study: The Context of Prior Findings on Coaching for the SAT

The Scholastic Aptitude Test (SAT) was developed as a measure of academic abilities, to be used toward the end of secondary school as a predictor of academic performance in college, and as a standardized supplement to the secondary school record available to college admissions officers. The SAT was explicitly designed to differ from achievement tests in school subjects in the sense that its content is drawn from a wide variety of substantive areas, not tied to a particular course of study, curriculum, or program. Moreover, it taps intellectual processes of comprehension and reasoning that may be influenced by experiences outside as well as inside the classroom. These intellectual skills are exercised to some degree in all subject-matter areas at all levels of schooling, as well as in response to real-life situations. These skills are thereby learned in manifold ways and gradually develop over time as a function of both instruction and experience. The growth of these acquired abilities thus tends to be slower than growth in achievement areas such as French, chemistry, ancient history, and trigonometry, which are primarily learned as a function of curriculum and instruction. In addition, the SAT is intended as a prospective measure for the prediction of future academic performance rather than as a retrospective measure for the evaluation of past school-related attainment. The specific item content on the SAT attempts to sample the sort of cognitive skills underlying college-level performance: reading with comprehension, understanding vocabulary, verbal reasoning, computational skills, quantitative reasoning, and problem solving.

Thus, the purpose of the SAT is to predict academic success in college. It was designed to reflect verbal and mathematical abilities that are acquired over an extended period and hence can be expected to be difficult to improve substantially through short-term instructional efforts. However, these reasoning abilities are learned and if effective learning experiences, including effective coaching programs, facilitate their development, one would expect not only increases in SAT scores but also increases in crite-
rion performance—in this case, performance in college. Such increases in test scores and academic performance do not invalidate the SAT—rather, they contribute to its construct validity and predictive utility. Coaching or special preparation programs that increase test scores that were inaccurately low because of anxiety, for example, or test unfamiliarity would also improve the validity of measurement by virtue of reducing the prior invalidity of measurement. If, on the other hand, special preparation or coaching efforts result in a substantial increase in test scores without a corresponding positive effect on the level of verbal and mathematical reasoning abilities measured by the test, the value of the test scores for admissions purposes would be undermined.

Thus, three major potential effects of coaching may be distinguished: First, some coaching programs may improve the abilities and skills measured by the test, resulting in commensurate increases in test scores; second, some coaching programs may enhance test-taking sophistication or reduce anxiety associated with taking tests, resulting in increased test scores that are more accurate assessments of ability and skill; and, third, some coaching programs may teach test-taking strategies and answer-selection tricks, resulting in increased test scores that are inaccurately high. Some coaching programs, of course, may produce none of these effects or more than one in various combinations. The first two effects, if they were realized, would be good from the standpoint of the student and good from the standpoint of test validity. The third effect is probably minor with well-constructed tests, because test-makers should strive to minimize the use of complicated or tricky item formats and to eliminate items that may be answered on the basis of clues unrelated to the abilities tested. Moreover, test-makers should also strive to reduce the import of the second possible effect of coaching by providing test-familiarization materials and practice tests to all candidates as well as advice on guessing, reviewing, pacing, and the like, in part to reduce apprehension about what to expect. With respect to the first type of effect, if coaching or special preparation programs can improve the abilities measured by the SAT, the effective components and techniques should be identified and incorporated more widely in secondary- and possibly elementary-school instruction. There is little question that the verbal and mathematical abilities measured by the SAT are learned, but there are large questions about how they can be taught.

There is a progression or ordering of educational tests ranging from measures of scholastic and intellectual abilities at one pole,
with content drawn from a variety of substantive areas, to measures of academic attainment at the other pole, with content specialized by subject-matter field. The SAT falls toward the first extreme—it taps general intellectual processes that develop gradually over many years of experience and use both inside the classroom and in everyday life, and these processes should therefore be relatively difficult to enhance markedly through brief courses of intervention. The typical educational achievement test falls toward the other extreme—it taps specific knowledge and skills acquired through the normal course of classroom instruction or independent study, and this knowledge and skill should therefore be relatively responsive to instructional intervention, even in brief courses. Similarly, there is a progression of types of preparation for taking examinations ranging from simple practice on sample items at one extreme to the provision of intensive instruction aimed at developing ability and knowledge at the other extreme. What has come to be called "coaching" is now generally considered to fall anywhere in the broad range between these two extremes, entailing some combination of test familiarization, drill-and-practice with feedback, training in strategies for general test-taking and for specific item formats, subject-matter review, and skill-development exercises—although widespread usage of the term in the past tended to underscore drill and cramming toward the practice end of this range. The important point, however, whatever the special test preparation is called, is whether or not it leads to significant test score increases and, if it does, whether those increases represent genuine long-term improvements in the knowledge and abilities measured by the test as opposed to enhanced test-taking skills irrelevant to the purposes of testing. The studies that have been undertaken in this area, including the 

Although there was considerable variation in the score increases observed for particular groups of students and for particular programs or treatments, the earlier studies on the average reported score increases associated with special preparation, relative to score increases for control groups, of less than 10 points on SAT-Verbal and less than 15 points on SAT-Math, on a score scale ranging from 200 to 800 points. Studies lacking control groups yielded larger effects but since they differed from control-group
studies not only in design characteristics but in critical program characteristics, their interpretation is especially problematic. In spite of an apparent consistency of results when some type of control-group comparison was included, many of these studies had methodological weaknesses that detract from the strength of the conclusions drawn. We now summarize the results of these earlier studies and consider in detail their particular strengths and weaknesses. The FTC study will be analyzed in the next section of this report.

**Studies With No Control Groups**

Three studies have been conducted—those by Pallone in 1961, by Marron in 1965, and by Coffman and Parry in 1967—that lacked any control group for evaluating unusual patterns of score change. Pallone (1961) looked at the effects of short- and long-term intensive developmental reading courses on SAT scores of students in a private school for boys. The courses were undertaken "for students in their final year of pre-college work, including a large number of high school graduates who were completing a year of post-high school study in preparation for entrance into the U.S. government academies" (p. 655). According to Pallone, to improve the skills measured by the SAT, "not 'coaching' methods, but instruction of a developmental nature in reading and vocabulary skills was indicated. Improvement in scores could be expected only if the basic skills measured by the test were first strengthened" (pp. 654-655). This program provided focused instruction to strengthen reading achievement along with intensive practice in reading skills, including such special skills as skimming and critical reading, as well as a brief analysis of typical verbal analogy test items. Approximately 20 students participated in a six-week summer pilot program that met for 90 minutes daily. Substantial percentile increases corresponding to mean reading achievement were reported on the basis of pre- and posttest results. An average score increase of 98 points was obtained on SAT-Verbal from the March 1959 to August 1959 administrations. The long-term program covered a six-month period from September to March with daily meetings of 50 minutes each. From March of 1959 to March of 1960 for some 80 students who completed the long-term course, an average SAT-V increase of 109 points was reported, although the difference in mean scores in Pallone's Table 3 (p. 656) was only 84. The students who par-
ticipated in the summer program also completed the long-term reading course. The mean increase in their Verbal scores for the period from March 1959 to March 1960 was almost 122 points, or an average of about 24 points over the increase reported after the summer course. The special quality of the sample and the lack of control groups severely limit the implications of these findings vis-à-vis coaching. Furthermore, the instructional focus on skill development and the intensive and long-term nature of the programs put Pallone’s efforts close to what ordinarily would be considered “instruction” in contradistinction to “coaching,” as Pallone himself insisted.

In the absence of control groups of similar students at this preparatory school who were not taking Pallone’s course, it is difficult to assess the import of these score gains. Pallone [1961] suggested comparing them with normal expectations of gains of about 35 points on SAT-V during the final secondary school year, which would yield an instructional or program effect of about 83 points for the summer pilot program (prorated for the five months between tests) and about 74 points for the long-term program. This is not a very satisfactory comparison, however, since Pallone’s students were not representative of students in their final year of high school who take the SAT. Slack and Porter [1980] suggested comparing Pallone’s results with average gains in national administrations of junior- to senior-year retesters having the same initial average score levels as Pallone’s students, which yields prorated instructional effects of 85 and 79 points, respectively, for the summer and long-term programs. Again, this is not a very satisfactory comparison because Pallone’s private school students were not a representative sample of the national population of test repeaters. Pike [1978] suggested comparing Pallone’s results with average gains of control students in superior schools from other studies of proprietary programs—for example, those conducted by Frankel [1960] and by Whitla [1962] described below. When adjusted for time differences between pre- and post-tests in these studies, this comparison leads to instructional effects of about 75 and 53 points, respectively, for the summer and long-term programs. If compared with score gains of control students in other coaching studies who had average initial score levels roughly comparable to Pallone’s groups, namely in this instance a selection of control schools from the study by Dear [1958] described below, the adjusted instructional effects are 80 and 65 points, respectively. The point is that in the absence of comparable control groups, no generally satisfactory estimate of
instructional effects can be obtained. If an average is taken across the four adjustments just suggested, the resulting estimates of instructional effects are 81 and 68 points, respectively, for the summer and long-term programs. These values are very likely still overestimates, however, because none of the comparisons can take into account the highly self-selected nature of Pallone's private school students, many of whom were completing a post-high school year highly motivated to increase their chances of entering service academies or selective colleges. But given the overall size of the effects, even with somewhat larger adjustments, it seems likely that Pallone's [1961] intensive summer and long-term efforts at "instruction of a developmental nature" may have succeeded to some degree in strengthening basic skills measured by the SAT.

Marron [1965] examined SAT score gains for students at ten well-known preparatory schools that specialize in preparing high school graduates for admission to the service academies and selective colleges. The instructional programs entailed "six months of full-time exposure to course content that is directly related to the verbal and mathematics College Board tests [both Aptitude and Achievement]" [p. 1]. A special administration of SAT-V and -M and College Board achievement tests (English composition and intermediate or advanced math) at all ten schools in September 1962 served as the pretest, while the posttest was the regular College Board admissions testing administration in March 1963. It should be noted that if the level of motivation and effort on a special pretest that did not count for college admission was not comparable to that on the regular posttest, the instructional effects in Marron's study would likely be overestimated. Since significant differences were obtained among the ten schools with respect to both the September pretest scores and the March posttest scores and these latter differences remained significant in analyses of covariance adjusting for pretest levels, the overall results were reported separately for groups of schools having non-significant differences within group. Score gains on SAT-V were 77 points for group 1 (2 schools, N = 83), 56 points for group 2 (6 schools, N = 600), 47 points for group 3 (1 school, N = 5), and 35 points for group 4 (1 school, N = 26); the weighted average SAT-V increase over all groups was 58 points. Score increases for SAT-M were 83 points for group 1 (4 schools, N = 232), 78 points for group 2 (3 schools, N = 405), and 72 points for group 3 (3 schools, N = 78); the weighted average SAT-M increase over all groups was 79 points. Weighted average gains on achievement tests were 83
points for English composition, 133 points for intermediate math, and 127 points for advanced math. Again, in the absence of control groups, it is difficult to appraise the size of these instructional effects, but it appears that on a relative basis achievement test scores improved much more than SAT scores—on the average about half again as much.

Marron [1965] suggested comparing the SAT score gains with those considered typical for males in their senior year in secondary school, which he reported based on College Board data as 40 points for SAT-V and 43 points for SAT-M over a 10-month testing interval. Prorated for the six-month testing interval in Marron's study, this yields an adjusted weighted average of 34 points for SAT-V and 53 points for SAT-M. As was done for Pallone's [1961] results, the suggested adjustments of Slack and Porter [1980] and Pike [1978] were also applied to Marron's figures, along with an adjustment based on score gains of control students in other coaching studies who had average initial score levels roughly comparable to Marron's groups. If an average of all four of these adjustments is taken, the resulting weighted average values are 35 points for SAT-V and 54 points for SAT-M. But again, none of these adjustments is very satisfactory because none of the suggested comparisons takes into account the highly self-selected nature of Marron's students, thereby leaving important factors of differential motivation and growth uncontrolled. In any event, the relevance of Marron's [1965] study to the issue of 'coaching' is arguable, since six months of full-time exposure to course content directed at verbal and mathematical knowledge and skills would ordinarily be considered 'instruction.' These score gains, whatever comparison groups they are contrasted with, might better be interpreted as testimony to the learning and skill development that highly motivated high school graduates can accomplish with six months of full-time concentrated effort on focussed curricula.

In an attempt to explore further the effects of developmental reading instruction on SAT-V scores, especially in light of Pallone's [1961] findings, Coffman and Parry [1967] undertook a study of three groups of college freshmen who took the SAT-V before and after completing a course in accelerated reading. The course was described as stressing speed of reading with relative accuracy. Pre- and posttest scores based on special administrations of the SAT were available for two small groups of 10 and 9 students who elected an eight-week course meeting six hours each week. Pre- and posttest SAT-V scores were also available for
25 students whose course met three hours a week for 15 weeks. For the eight-week course, the increases in SAT-V were 3.5 and 9.9 points, while a 28.9 mean loss was observed for the group taking the 15-week course. This score decrease in the 15-week course may stem from problems of test administration and score equating since time constraints dictated the use of a shortened SAT-V for that group. It may also reflect problems in the motivation of students who take a special SAT when they are already in college, as may the relatively modest score increases in the other two groups—although all of these students were presumably motivated to enroll in the course, which explicitly entailed taking the SAT. In addition, the lack of control groups is again a factor that seriously impairs the usefulness of these results. Although these findings are in sharp contrast to the score increases reported by Pallone, the accelerated reading program studied by Coffman and Parry appears to be considerably different from Pallone’s developmental reading curriculum and apparently not as directly relevant to the skills measured by the SAT.

**Studies With Nonequivalent Control Groups**

A second methodological weakness obtains in three other earlier studies: Although each incorporated control groups, these groups were drawn from schools different from those providing special preparation for the experimental students, thus confounding treatment or preparation effects with school effects.

Dyer (1953) studied seniors at two highly selective independent schools for boys—225 students at one school served as the treatment group and 193 students at the other school served as the control group. The students at both schools took the SAT twice, once in September 1951 and again in March 1952. The two groups of students were similar with respect to length of time enrolled in the school, level of SAT scores, and the level and number of foreign-language and math courses taken. The experimental group completed 12 verbal practice exercises in 30- to 60-minute sessions and five math practice exercises in 60- to 90-minute sessions. The control variables in an analysis of covariance were initial SAT scores, number of years each student had been enrolled in school, and the number and level of foreign-language and mathematics courses taken in the senior year. The estimated increase in score for the treatment group over the control group was about 4½ (4.6) points in Verbal and about 13 (12.9) points in Math.
When the students were divided into those who were not taking mathematics courses as seniors and those who were, the no-math boys who were coached gained over 29 points more than those who were not coached. In contrast, it was found that the boys taking mathematics who were coached gained 3.3 points more than those taking mathematics who were not coached. Dyer's conclusions were [1] that coaching on sat-Verbal is not likely to be effective; [2] coaching on sat-Math will be of some advantage only if the students coached are not already enrolled in regular math courses.

In 1955, John W. French reported a study conducted at three schools: The students at School A (N = 158) pursued their regular courses with no attempt at special preparation for the sat and served as a control group for coaching in both Verbal and Math; the students at School B (N = 110) served as a treatment group for Verbal and as a control group for Math; and, the students at School C (N = 161) served as a treatment group for both Verbal and Math. The special preparation program in Verbal at School B differed from the one at School C in that the former primarily emphasized vocabulary for a total of 4½ hours, while the latter reviewed 10 verbal exercises more representative of sat-V skills. The students were all seniors who planned to go to college. In addition, some students in School C participated who were enrolled in a technical course. All students took the sat in September 1953 and again in March 1954. The data were subjected to an analysis of covariance, using sat scores earned in September 1953 as the independent or control variable and sat scores earned in March 1954 as the dependent or outcome variable. The results of French's analysis were mixed. The advantage in Verbal score for boys and girls combined who were coached over those who were not coached was found to be 18 points in one school and 5 points in another (the one with the vocabulary coaching program). The score increase in Math for boys and girls combined who were coached over those who were not coached was shown to be 6 points when compared to one control school and 18 points when compared to the other. Thus, the largest benefit resulting from coaching for groups of boys and girls combined was 18 points in Verbal and 18 points in Math. When broken down by sex and current enrollment in math courses, the sat-M data indicated that coached boys not studying math at the time of coaching showed greater increases over control students than did coached boys who were studying math, by about 4 points in one school and 10 in the other. This pattern of higher coaching effects for boys not
currently studying math is consistent with the Dyer results. In contrast, coached girls not studying math at the time showed smaller increases over control students than did coached girls who were studying math. Coached girls currently studying math exhibited score increases of about 20 and 30 points over the non-coached girls currently studying math in the two control schools, whereas the coached girls not studying math exceeded their non-coached counterparts in the two math control schools by only 1 and 4 points.

French also investigated the effects of using some identical items in coaching and on the posttest. Shortly before the March SAT administration, the students in School C were given a practice test containing 50 verbal and reading items and 34 math items, half of which had been included in the previous practice exercises of the coaching course. Students at a fourth school who received no special preparation served as a control group. The raw scores on the familiar and unfamiliar halves of the test were each converted to the SAT 200- to 800-point scale, and in an analysis of covariance the score on the coached items was predicted from the score on the uncoached items. A comparison of the results showed an increase of 15 points on SAT-M and 47 points on SAT-V for the coached over the uncoached students on the projected full-length test made up of items explicitly coached. These findings relate to a test made up of items identical to those on which students received coaching, and it is unlikely that this situation would ever arise in practice. Nevertheless, in 1974 June Stern included an identical subset of items in the April SAT administration and again in the November administration with no experimental coaching intervention. She found that the average net gain in Verbal as well as in the Math score was approximately 17 points for repeaters of identical items over gains for repeaters of non-identical items. Clearly, if just taking an identical subset of items twice yields a 17-point increase, the effects of coaching on identical items, if one had access to them, could be substantial.

In 1958, Robert Dear reported a study to determine whether longer periods of coaching in small groups—two class periods a week for 6 weeks and for 12 weeks—were likely to be more effective than the shorter-term coaching studied by Dyer and by French. In Dear’s study, 6 public and 4 private secondary schools were chosen randomly from a list of schools from which at least 15 students had taken the SAT as juniors in May of 1956. A treatment group from each school was selected at random from students who volunteered for coaching. A second group of nine
schools from the same geographical region—the New York-New Jersey-Greater Philadelphia area—was drawn at random as the group of control schools. All students eligible for the study had indicated an interest in being coached. Three students were selected from each school from each of three ability levels—90 coached students and 81 control students. Of these, 71 coached and 79 uncoached students took the SAT in May 1956 and in March 1957. The coaching program began in mid-November and continued through mid-March with weekly coaching sessions supplemented by one additional hour of homework each week. Most students repeated the SAT in January 1957, halfway through the coaching period, and again in March 1957. The January results showed about a 22-point advantage for 60 coached students on Math, but a 2 1/2-point disadvantage on Verbal. The control groups for this January administration included all uncoached students (N = 526) in both the same schools and the control schools who indicated interest in being coached. The March advantage for the coached students on Math was about 24 points. The Verbal score results, unfortunately, were not determined because of a significant difference in the slopes of the regression lines for the coached and uncoached groups.

Studies With Matched Control Groups

A third methodological problem occurs in studies in which control and experimental students, although from the same school, are not assigned randomly but are matched on selected measures, thereby still permitting systematic differences between the groups on other nonassessed variables. In 1960, Edward Frankel published the first study on the effects of commercial coaching on SAT scores. Frankel selected 45 high school students who had taken commercial coaching courses and matched them with 45 control students from the same high school in the following ways: (1) their pretest scores on the May 1958 SAT were within 10 points of each other on both Verbal and Math; (2) they had taken the same form of the SAT in either December 1958 or January 1959 to serve as a posttest; and, (3) they were of the same sex. Within each pair, one had taken a commercial coaching course involving roughly 30 hours of coaching in classes of about 25 students between May 1958 and January 1959. In brief, Frankel found an 8.4 point advantage for coached students on Verbal and a 9.4 point advantage on Math.
In 1962, Dean Whitla compared the score increases of 52 students who had attended an intensive 10-hour course in improved study habits, reading skills, and math concepts at a proprietary school in Boston with the score increases of a comparable group of 52 students from the same area who had not taken a coaching course. All of the students had taken the SAT in March or May of their junior year and, in addition, were administered a special SAT in the fall of 1959 when the study began; the January 1960 SAT served as the posttest. The average V and M scores of the two groups were within one point of each other on the spring SAT and within two points on the fall pretest, suggesting that the two groups were not only well-matched in terms of initial level but also in terms of growth rate over this period. Whitla found an 11-point advantage for the coached group on Verbal but a 5-point disadvantage on Math when the posttest was compared with the fall pretest; there was a 10-point advantage on Verbal and a 7-point disadvantage on Math when compared with the spring SAT.

Studies With Randomized Control Groups

Three studies employing a randomized design have been carried out. The first one, by Roberts and Oppenheim (1966), utilized the Preliminary Scholastic Aptitude Test (PSAT) as both pretest and posttest. In contrast to earlier coaching studies that involved highly selective and effective private schools and specialized or suburban public schools, the Roberts and Oppenheim study was undertaken to investigate whether students receiving less adequate instruction might especially benefit from special preparation. Data were collected from 18 predominantly, if not entirely, black secondary schools in rural and urban Tennessee. In 6 schools coaching consisted of special instruction in verbal material, in 8 schools coaching was for mathematics, and in 4 schools no special instruction was provided. Within the treatment schools students were assigned randomly to coached and uncoached groups. The instruction was provided in 15 half-hour sessions over a 4- to 6-week period. The results showed small increases for the coached groups over the control groups: about 1½ (1.44) points on PSAT-Verbal and less than 1 (.81) point on PSAT-Math, increases which correspond to about 14 points on SAT-Verbal and 8 points on SAT-Math. This advantage of coached over control groups, however, was due as much to score decreases on the part of the control students, possibly signaling problems in
motivation or attrition, as to score increases on the part of the coached students.

A second randomized study, conducted by Evans and Pike (1973; Pike & Evans, 1972), examined intensive coaching efforts in the math area. A sample of 509 students in 12 schools participated in the study. The coached students received 21 hours of instruction and 21 hours of homework, over a 7-week period during November and December 1970, directed at one of the following item types: Regular Math [RM], Data Sufficiency [DS], or Quantitative Comparisons [QC]. Three randomly chosen groups of students were defined in each school: one to be instructed in QC, one in either RM or DS, and one as a control group. These groups took the SAT first in October 1970, which served as a pretest, and again in December 1970 [posttest] and April 1971 [delayed posttest]. The pretest and posttest were special administrations of the SAT, whereas the delayed posttest was a regular administration. The three experimental groups were given special preparation during November and December, and the control group received instruction after the December posttest. On this schedule, all groups received instruction in test-taking skills—becoming familiar with test directions, pacing, and appropriate strategies for using partial information and guessing. All groups also had instruction in math content—numerical facts, numerical and basic algebraic skills, and in particular mathematical areas such as inequalities. In addition, there was practice on one of the item types for students in each respective experimental group. The study revealed score increases beyond those experienced by the control group for each of the three experimental groups coached on a particular item type. However, because the Evans-Pike study was designed to investigate the relative susceptibility of three item formats to special instruction, it is difficult to say just how large the effects were in terms of SAT scores. The authors’ best estimate of score increases reflective of coaching for all four groups over the total period from October to April was about 25 points. The average increase over the control group for the three experimental item-type groups, weighted according to their respective sample sizes, was 16.5 points over the period from the October pretest to the December posttest. Pike (1978) later conjectured that, still keeping within the total 21 hours of special preparation, a judicious combination of instruction for both RM and DS, the two major item-types then in SAT-M, would be expected to yield coaching or special preparation effects of about 33 points.

In the third randomized study, Alderman and Powers (1979) in-
vestigated the effectiveness of existing secondary school programs that had been initiated by the schools to improve the performance of students on the SAT-Verbal scale. Students at each of eight schools for whom PSAT scores were available were randomly assigned to a special-preparation group or to a control group. Access to the same preparation course was delayed for control students for the purpose of this study. A special administration of a retired SAT was used as the posttest. Across the eight schools the overall increase in SAT-V attributable to special preparation was about eight points, which is statistically significant at the .05 level. The actual effects ranged from -3 points at one school to +28 at another. Differences in effectiveness between the coaching programs were not statistically significant from school to school, however, and the best statistical estimate of the range was from 4 to 16 points. Nor, apparently, did the control groups react in comparable fashion from school to school, possibly reflecting differences in motivation or seriousness in approaching a special SAT. The largest school effect of 28 points, for example, resulted in part from a control-group decrease almost equal in magnitude to the treatment-group increase (-11 score points in going from the converted PSAT to the SAT for the control group versus +13 score points for the treatment group, which yields a 28-point school effect after covariance adjustments are made). An attenuated form of this pattern occurred at two other schools, whereas the remaining five schools showed varying degrees of score increase for both treatment and control groups. Overall, only one-tenth of one percent of the total variance in SAT-V scores was accounted for by special preparation alone, while school attended taken alone accounted for 24 percent of the Verbal score variance. Furthermore, scores on the PSAT-V and the Test of Standard Written English accounted for 54 and 39 percent of SAT-Verbal score variance, respectively. The authors concluded that the school attended and scores on earlier tests of verbal and writing ability contribute much more to determining SAT-V scores than did special preparation as typically offered by secondary schools.

Another, more subtle methodological problem has emerged in the process of reviewing these coaching studies having randomized control groups, namely the problem of engendering and maintaining realistic motivation for taking the posttest SAT, especially for uncoached control students. Developing realistic motivation and effort for taking pre- and posttests is a common requirement of all coaching studies. However, the three experiments that employed randomized control groups happened also
to use as the posttest a special administration of the SAT or PSAT rather than a regular administration. These special administrations may have been viewed to some degree as practice tests that did not count for the record, thereby eliciting less motivation and effort than would a regular SAT administration. Warning signals suggesting this possibility were noted in the Alderman and Powers (1979) study, in which control groups in three schools were found to exhibit score decreases in going from a regular PSAT to a special SAT, even though the expectation is for a score increase from an October PSAT to a Spring SAT of upwards of 10 or 12 points, which is the basal estimate provided by national administration samples. Such control-group score decreases may substantially complicate the interpretation of the results. One way of viewing the Alderman and Powers analysis, for example, is that their appraisal of the import of within-school intercept differences tacitly implies that within each of the eight schools the effect on student performance of taking a special rather than a regular SAT operates consistently for both uncoached and coached students on the average. Differences in performance between these groups at a given school may then be used to estimate coaching effects on the implicit assumption that any reduced motivation and effort on the special SAT posttest would have lowered mean performance levels of both groups to about the same degree. Such assumptions are not as apparent when the analysis is viewed in terms of sampling fluctuations across eight independent samples: The control-group score changes from pretest to posttest ranged from -11 to +30 with a mean of 8, and the treatment-group score changes ranged from 3 to 44 with a mean of 17; the 9 point overall advantage for the coached over the control students, when adjustments for covariates are made, corresponds to the 8 point coaching effect reported in the study.

Control-group scores also decreased in the Roberts and Oppenheim (1966) study, which similarly employed a special administration posttest, in that case a special PSAT. Unlike Alderman and Powers, who used a regular administration of the PSAT as a pretest, the Roberts and Oppenheim pretest PSAT was a special administration as well. None of the nonrandomized coaching studies reviewed earlier displayed control-group score decreases, and the posttest in all of those studies was a regular administration of the SAT—although Dyer (1953), French (1955), and Whita (1962) did use special pretest administrations and thereby introduced the possibility of other biases. The remaining randomized study, conducted by Evans and Pike (1973), employed special adminis-
tration SATs as both pretest and posttest, but their delayed posttest was a regular SAT administration. Pike's (1978) subsequent interpretation of score increases from posttest to delayed posttest as reflecting the long-term consolidation or continuance of gains due to coaching becomes jeopardized from this vantage point by the plausible rival interpretation that those score gains instead reflect increases in motivation and effort in going from a special administration to a regular one.

**Comparison of Results Across Studies of Coaching**

For numerous reasons, including the diversity of design limitations and the differences in sample sizes, it is difficult to compare results across these several studies in a meaningful way (cf. Pike, 1978). Table II-1 represents one such attempt for those studies having some type of control group. The size of coaching effects reported there were calculated uniformly as follows: When analysis of covariance was performed, the reported values are intercept differences between the experimental and control regression lines, weighted in the case of multiple experimental or control groups by their respective sample sizes. In four studies not reporting analyses of covariance, the values in Table II-1 are average score increases of experimental over control groups, again weighted in the case of multiple experimental or control groups by their respective sample sizes. Two of these latter studies (by Frankel and by Whitta) involved statistical matching, and two (by Roberts and Oppenheimer and by Evans and Pike) involved randomization. Averaging these results over all of the studies in Table II-1, weighting in each case by the size of the experimental sample, yields 9.1 points for Verbal and 13.0 points for Math (the unweighted averages are 8.5 points for Verbal and 12.3 points for Math).

For those studies having no control groups, a summary is provided in Table II-2. The special preparation programs summarized there focus on verbal or mathematical content knowledge and skill development and entail the largest amounts of student contact time of any of the studies reviewed, which would ordinarily lead one to characterize them as instruction rather than coaching. These programs range from 45 hours of student contact time over 6 weeks to 48 hours over 8 weeks to roughly 100 hours over 6 months to virtually full-time over 6 months or approximately 600 hours. In contrast, the most intensive of the control-group stud-
## Table II-1
Average Difference Between Experimental and Control Groups in Studies of SAT Interventions [Adapted from Donlon]

<table>
<thead>
<tr>
<th>Study/Design</th>
<th>Sample Characteristics</th>
<th>Characteristics of the Special Preparation</th>
<th>SAT-Verbal Difference(^1)</th>
<th>Significance Level(^2)</th>
<th>N / Control</th>
<th>SAT-Math Difference(^3)</th>
<th>Significance Level(^2)</th>
<th>N / Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyer [1953]</td>
<td>Private H.S. Seniors M</td>
<td>Twelve 30-60 minute sessions for verbal; five 60-90 minute sessions for math.</td>
<td>4.6</td>
<td>.05</td>
<td>225/193</td>
<td>12.9</td>
<td>.01</td>
<td>225/193</td>
</tr>
<tr>
<td>Control, different school</td>
<td>Public H.S. Seniors M + F</td>
<td>Ten verbal and ten math coaching sessions using ETS item materials.</td>
<td>18.3</td>
<td>.01</td>
<td>161/158</td>
<td>6.2</td>
<td>.01</td>
<td>161/158</td>
</tr>
<tr>
<td>French [1955]</td>
<td>Public H.S. Seniors M + F</td>
<td>Total verbal coaching 4½ hours; math coaching was ten sessions using ETS item materials.</td>
<td>5.0</td>
<td>.05</td>
<td>110/158</td>
<td>18.0</td>
<td>.01</td>
<td>161/110</td>
</tr>
<tr>
<td>Control, different school</td>
<td>Public &amp; Private H.S. Seniors M + F</td>
<td>Approximately 6 weekly 2-hour, 2-person coaching sessions, plus 1 hour of homework each week.</td>
<td>-2.5</td>
<td>N.S.</td>
<td>60/526</td>
<td>21.5</td>
<td>.01</td>
<td>60/526</td>
</tr>
<tr>
<td>Dear [1958]</td>
<td>Public &amp; Private H.S. Seniors M + F</td>
<td>Approximately 12 weekly 2-hour, 2-person coaching sessions plus 1 hour of homework each week.</td>
<td>-3.0</td>
<td>N.S.</td>
<td>71/116</td>
<td>23.6</td>
<td>.01</td>
<td>71/116</td>
</tr>
<tr>
<td>Control, same and different schools</td>
<td>Public &amp; Private H.S. Seniors M + F</td>
<td>Ten 3-hour, 25-person sessions of coaching.</td>
<td>8.4</td>
<td>N.S.</td>
<td>45/45</td>
<td>9.4</td>
<td>N.S.</td>
<td>45/45</td>
</tr>
<tr>
<td>Frankel [1960]</td>
<td>Public H.S. Seniors M + F</td>
<td>Commercial proprietary school. Five 2-hour sessions plus intensive homework. Verbal and math.</td>
<td>11.0</td>
<td>N.S.</td>
<td>50/50</td>
<td>-5.3</td>
<td>N.S.</td>
<td>50/50</td>
</tr>
<tr>
<td>Control, statistically matched</td>
<td>Public &amp; Private H.S. Seniors M + F</td>
<td>7½ hours of programmed instruction in test-taking and in verbal and math content.</td>
<td>14.4(^*)</td>
<td>.05</td>
<td>188/122</td>
<td>8.1(^*)</td>
<td>N.S.</td>
<td>288/129</td>
</tr>
<tr>
<td>Whita [1962]</td>
<td>Public &amp; Private H.S. Seniors M + F</td>
<td>No coaching for SAT-V</td>
<td>—</td>
<td>16.5</td>
<td>.05</td>
<td>239/320</td>
<td>No coaching for SAT-M</td>
<td>—</td>
</tr>
<tr>
<td>Control, statistically matched</td>
<td>Public &amp; Private H.S. Seniors M + F</td>
<td>Varied strategies, at eight schools, centering on reading and analogies; time range 3-45 hours.</td>
<td>8.4</td>
<td>.05</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Average weighted by size of experimental sample.

| | 9.1 | 13.0 |

\(^1\)The coaching effects are intercept differences between regression lines for experimental and control groups or [for Frankel, Whita, Roberts & Oppenheim, and Pike & Evans] average score increases of experimental over control groups, both weighted in the case of multiple experimental or control groups by their respective sample sizes.

\(^2\)As shown for coaching effects reported in original text.

\(^3\)Not calculated; variances and regression slopes differed significantly for experimental and control groups.

\(^*\)This study employed the SAT as both pre- and posttest; the averages shown have been converted to the SAT score scale ranging from 200 to 800 points.
ies summarized in Table II-1 were 30 hours (plus homework) over 10 sessions (Frankel, 1960), 24 hours (plus 12 hours homework) over 12 weeks (Dear, 1958), and 21 hours (plus 21 hours homework) over 7 weeks (Evans & Pike, 1973).

In the absence of control groups, instructional or program effects were estimated in Table II-2 in the manner described earlier—that is, by adjusting the average score gains reported in each study by the average of four adjustments, those suggested by (1) the authors of the original articles, (2) Slack and Porter (1980), and (3) Pike (1978), as well as (4) the average score gains of control students in other coaching studies who had roughly comparable initial score levels. Averaging these estimates over all the studies in Table II-2, weighting in each case according to group size, yields 38 points for sat-V and 54 points for sat-M (the unweighted averages are 39 points for Verbal and 53 points for Math). Given the dubious and provisional nature of the adjustments and the highly self-selected character of the students in each program, these values are still probably overestimates of program effects. But their general magnitude suggests that the verbal and mathematical reasoning skills measured by the sat may be enhanced to a measurable degree by long-term and intensive instruction, at least for highly motivated students.

Granted that there is some overlap or blurring of the distinction between coaching studies in Table II-1 and instructional studies in Table II-2, the two types seemed sufficiently different to warrant separate treatment. Accordingly, overall averages were not computed for the total combined set of studies because possible differences in impact might thereby be obscured. In contrast, Slack and Porter (1980) have chosen to combine both types of studies in a single table and to report overall weighted average score increases of coached groups over uncoached control groups (or norm comparison groups); these weighted averages were 29 points for sat-V and 33 points for sat-M. However, their table included French’s (1955) study of identical items, which hardly seems appropriate, and a study by Lass (1958) which we viewed as too informal to treat quantitatively, and it did not include the studies by Alderman and Powers (1979) or by Evans and Pike (1973). If the French identical-items study is deleted from their table and the two missing studies are added and some corrections are made in the effective sample sizes of the studies by Dyer (1955) and Pallone (1961), the revised weighted averages are 25 points for Verbal and 33 points for Math. Further, if the score increases taken from the studies by Pallone (1961) and Marron
<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Characteristics</th>
<th>Characteristics of the Special Preparation</th>
<th>SAT-Verbal Adjusted Average Score Increase</th>
<th>N</th>
<th>SAT-Math Adjusted Average Score Increase</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pallone [1961]</td>
<td>Private</td>
<td>H.S. Seniors &amp; Graduates</td>
<td>90-minute daily instruction and practice in developmental reading skills over 6 weeks</td>
<td>81</td>
<td>20+</td>
<td></td>
</tr>
<tr>
<td>Pallone [1961]</td>
<td>Private</td>
<td>H.S. Seniors &amp; Graduates</td>
<td>50-minute daily instruction and practice in developmental reading skills, with stress on logical inference and analogic analysis over 6 months</td>
<td>68</td>
<td>80-</td>
<td></td>
</tr>
<tr>
<td>Marron [1965]</td>
<td>Private</td>
<td>H.S. Seniors &amp; Graduates</td>
<td>Full-time daily sessions aimed at verbal and math content and test-taking skills over 6 months</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coffman &amp; Parry [1967]$^2$</td>
<td>Public College Freshmen</td>
<td>M + F</td>
<td>6-hours weekly instruction in accelerated reading over 8 weeks</td>
<td>4$^1$</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

Weighted Average: 38 54

---

$^1$To estimate instructional or program effects, average score increases in the Pallone [1961] and Marron [1965] studies were adjusted by the average of four adjustments: those suggested by [1] the authors of the original articles—Pallone suggested 35 points on SAT-V as normal expectation of gains during the final year of secondary school and 15 points for the 5-month interval between tests in the short-term program, and Marron suggested 24 and 26 points, respectively, for SAT-V and SAT-M as typical gains for high school seniors over 6 months; [2] Slack and Porter (1980)—average gains in national administrations of junior- to senior-year retesters having the same initial average score levels as Pallone's and Marron's groups; and [3] Pike (1978)—average gains of control students in superior schools from other studies of proprietary programs; as well as [4] average gains of control students in other studies who have average initial score levels roughly comparable to Pallone's and Marron's groups.

$^2$The 15-week program in Coffman and Parry [1967] was not included because the 29-point mean decrease in scores was considered atypical and possibly indicative of motivational and test administration problems.

The two 8-week programs in Coffman and Parry [1967] were combined, but adjustments were made only by the Slack and Porter (1980) procedure, which attenuated by only a few points an already tenuous effect. None of the suggested comparison groups of SAT takers appeared to provide even remotely reasonable yardsticks for gauging score gains of students already enrolled in a college not requiring the SAT.
are adjusted by the average of four suggested control contrasts, as was done in this report's Table II-2, the adjusted weighted averages of Slack and Porter become 22 points for Verbal and 28 points for Math.

Thus, the overall effect of averaging together the two types of studies—those with control groups and those without—is to raise the estimates of average coaching effects based on control-group studies alone (Table II-1) by about 13 points on Verbal and 15 points on Math. This assumes that the adjustments applied to the score gains in studies lacking control groups were large enough to correct appropriately for the experiential growth of self-selected students that would have occurred regardless of the program. This combined averaging of Slack and Porter (1980) is misleading not only because the adjustments are questionable, but because the combined averages obscure important differences between the special preparation programs in the two types of studies. As indicated earlier, a comparison of the brief program descriptions in Tables II-1 and II-2 reveals that the programs in studies lacking control groups happened also to be quite long-term and intensive with respect to student contact time, while the programs in control-group studies were relatively short-term and nonintensive. The former programs also entailed organized curriculum content and skill development as well as test review, whereas the latter programs tended to emphasize test review and practice exercises.

Rather than averaging across these studies, which inevitably precipitates arguments about the appropriate size of the score effects to be included from uncontrolled studies, let us instead rank the studies in order of the reported treatment vs. control group contrasts and when control groups are not available in order of the reported score increases. That is, for SAT-V Pallone's (1961) long-term and summer programs would be ranked 1 and 2, respectively, followed by Marron's (1965) four groups, etc. This procedure grants that the score effects in the studies lacking control groups are larger in an ordinal sense than those in the control-group studies, but it takes no position with respect to how much larger. If the programs are then also ranked in terms of the number of student contact hours involved and a Spearman rank-order correlation coefficient computed (Table II-3), the rank correlation is found to be .60 across 22 studies for SAT-V and .80 across 11 studies for SAT-M. Both coefficients are significant at the .01 level (Dixon & Massey, 1951). If five particularly suspect studies are deleted from the calculations for SAT-V, the new correlation is .73 across 17 studies, which is also significant at the .01 level. In this
### Table II-3

Correlations Between Rank Order of Score Effect and Rank Order of Student Contact Time for Studies of SAT Interventions With and Without Control Groups

<table>
<thead>
<tr>
<th>Study</th>
<th>Verbal 1</th>
<th>Math 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Student Contact Time</td>
<td>Rank Score Effect</td>
</tr>
<tr>
<td>Dyer (1953)</td>
<td>10 hours</td>
<td>11</td>
</tr>
<tr>
<td>French (1955)</td>
<td>8.3 hours</td>
<td>12</td>
</tr>
<tr>
<td>French (1955) Vocab</td>
<td>4.5 hours</td>
<td>17</td>
</tr>
<tr>
<td>Dear (1958)</td>
<td>6 hours</td>
<td>13.5</td>
</tr>
<tr>
<td>Frankel (1960)</td>
<td>15 hours</td>
<td>8</td>
</tr>
<tr>
<td>Whita (1962)</td>
<td>5 hours</td>
<td>15.5</td>
</tr>
<tr>
<td>Alderman &amp; Powers (1979)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School C</td>
<td>10.5 hours</td>
<td>10</td>
</tr>
<tr>
<td>School E</td>
<td>6 hours</td>
<td>13.5</td>
</tr>
<tr>
<td>School F</td>
<td>5 hours</td>
<td>15.5</td>
</tr>
<tr>
<td>School G</td>
<td>11 hours</td>
<td>9</td>
</tr>
<tr>
<td>School H</td>
<td>45 hours</td>
<td>6.5</td>
</tr>
<tr>
<td>Pallone (1961)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td>45 hours</td>
<td>6.5</td>
</tr>
<tr>
<td>Long</td>
<td>100 hours</td>
<td>5</td>
</tr>
<tr>
<td>Marron (1965)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td>300 hours</td>
<td>2.5</td>
</tr>
<tr>
<td>Group 2</td>
<td>300 hours</td>
<td>2.5</td>
</tr>
<tr>
<td>Group 3</td>
<td>300 hours</td>
<td>2.5</td>
</tr>
<tr>
<td>Group 4</td>
<td>300 hours</td>
<td>2.5</td>
</tr>
<tr>
<td>Rank-Order Correlation</td>
<td>.73 (17 studies)</td>
<td></td>
</tr>
<tr>
<td>Roberts &amp; Oppenheim (1966)*</td>
<td>3.8 hours</td>
<td></td>
</tr>
<tr>
<td>Alderman &amp; Powers (1979)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>School A</td>
<td>7 hours</td>
<td></td>
</tr>
<tr>
<td>School B</td>
<td>10 hours</td>
<td></td>
</tr>
<tr>
<td>School D</td>
<td>10 hours</td>
<td></td>
</tr>
<tr>
<td>Coffman &amp; Parry (1967)*</td>
<td>48 hours</td>
<td></td>
</tr>
<tr>
<td>Rank-Order Correlation</td>
<td>.60 (22 studies)</td>
<td></td>
</tr>
</tbody>
</table>

1 When only a total student contact time was available, it was assumed that half the time was devoted to Verbal and half to Math.

2 Each of 12 exercises was estimated to require a 50-minute class period.

3 Each of 10 exercises was estimated to require a 50-minute class period.

4 Schools A, B, and D were suspect because of control-group score decreases and hence were omitted in these calculations based on 17 studies; they are included below in calculations based on all 22 studies.

5 Roberts & Oppenheim (1966) was suspect because of control-group decreases and hence was omitted from the calculations for Verbal based on 17 studies and for Math based on 10 studies.

6 Coffman & Parry (1967) was suspect because of treatment-group score decreases and other indications of motivational problems and hence was omitted from the calculations based on 17 studies. The 25-week program was dropped altogether because of the treatment-group score decreases, and the two 8-week programs were combined, yielding a weighted average score effect of 6.5.

7 Each of 5 "double-length" exercises was estimated to require two 50-minute class periods.

8 French's (1955) Math coaching group was contrasted with two control groups, one in a school having no coaching and one in a school having vocabulary coaching; these two comparisons were combined, yielding a weighted average score effect of 11.
latter calculation, Schools A, B, and D from Alderman and Powers (1979) were eliminated because of control-group score decreases, as was the Roberts and Oppenheim (1966) study for the same reason, and the Coffman and Parry (1967) study was dropped because of treatment-group score decreases and other indications of the low relevance to the SAT of both the accelerated reading program studied and the samples of enrolled college students employed. If one suspects study, namely Roberts and Oppenheim (1966), is deleted from the calculations for SAT-M, the new correlation is .74, which is also significant at the .01 level. These rank-order correlations are summarized in Table II-3. It should be noted that although the various coaching programs required different and unknown amounts of homework, this rank-correlation procedure tacitly assumes that the amount of homework in each case was roughly proportional to the number of student contact hours, so that the overall orderings would not be markedly changed if homework were taken into account.

In interpreting these sizable monotonic relationships between student contact time and score effects, it must be remembered that these are rank-order correlations between average values of different samples or groups, and correlations between averages are typically much higher than correlations between individual differences within groups. Furthermore, these rank correlations are dominated by the relative consistencies between the two types of studies with respect to the ranking variables—that is, the control-group studies are all relatively low in both student contact time and score effects while the uncontrolled studies are all relatively high in both student contact time and score effects. It must also be remembered that the relatively high-contact programs entailed structured curricula emphasizing content knowledge and skill development while the relatively low-contact programs emphasized test review and practice. With this confounding of program characteristics in mind, it appears that increases in student contact time (possibly serving as a proxy for increasing curriculum emphases on content knowledge and skill development) are systematically associated on the average with increases in SAT scores. However, even though the time dimension is covered in only a fragmentary fashion by the available studies, when the magnitude of (adjusted) score effect is plotted against student contact time, the relationship appears not to be linear but approximately logarithmic. There is somewhat more variance or noise around the line of best fit relating score effect to log contact time for SAT-V than SAT-M, but only the Pallone (1961) results appear to
be particularly out of line in this logarithmic formulation. Con- siderably more research is needed, of course, with special prepar- ation programs entailing more than 40 hours of student con- tact time to fill in the gaps on the time line. But, if this suggested logarithmic relationship has substance, then each additional increase in SAT scores may require geometrically increasing amounts of student contact time and of all the curricular effort that contact time may be proxy for.

In summary, the average coaching effect across studies having some type of control group was less than 10 points for SAT-Verbal and less than 15 points for SAT-Math. The special preparation pro- grams in these control-group studies tended to be relatively short- term and relatively nonintensive in terms of student contact time, three of the longest and most intensive being 30 hours of student contact over ten sessions, 24 hours of student contact over 12 weeks, and 21 hours over seven weeks. For particular groups of students and particular coaching programs, the score increases ranged as high as about 18 points for Verbal and about 24 points for Math (or even somewhat higher if interactions between years of math taken and sex are considered). The larger score increases tended to be associated with the longer and more intensive of these relatively short-term programs, especially in Math. The average program effect across studies having no control groups could only be tentatively estimated, because there is no good way of taking into account the experiential growth of self- selected students in the absence of comparable control groups. The amount of their experiential growth and skill development could be considerable in this case, because all of the programs in the studies lacking control groups happened also to be relatively long-term. In addition, they were relatively intensive in terms of student contact time. The briefest of these programs was 45 hours of student contact over six weeks, and the longest was virtually full-time over six months. The provisional estimate of average program effects for these noncontrolled studies was 38 points for Verbal and 54 points for Math. Although the substantive content of the coaching programs was not systematically evaluated in any of these studies, overall the smaller coaching effects appear to be associated with short-term, relatively nonin- tensive practice and review, and the larger effects (which occur more for Math than for Verbal) appear to be associated with longer-term, high student-contact programs focusing on skill development.
Beyond the FTC Study: Critique of Assumptions and Inferences

This section draws together the major points made in several reviews undertaken at Educational Testing Service (ETS) of the Federal Trade Commission's (FTC) study of the effectiveness of commercial coaching schools in raising scores on the College Board Scholastic Aptitude Test (SAT). The separate reviews overlap somewhat in their criticisms, but they are reproduced in full in the Appendix since some additional points not covered in this summary are raised. The comments that follow are limited to the revised statistical analyses issued by the FTC Bureau of Consumer Protection (BCP) in March of 1979, undertaken in response to the several major data-analysis flaws observed in the memorandum issued previously by the FTC Boston Regional Office (1978).

Before proceeding to develop the critical arguments, we will first briefly describe the object of this critique, namely the FTC study of the effects of commercial coaching on the SAT. Students enrolled in two New York City area commercial coaching schools during the testing years 1974-75, 1975-76, and 1976-77 served as the experimental or treatment group, and a random sample of uncoached persons who took the SAT during the same three-year period in the same greater New York metropolitan area served as a control group. Data from a third coaching school were not analyzed because of its small sample size. Six subgroups were analyzed: (1) high school juniors taking the SAT for the first time in April 1975 (76 coached and 607 uncoached students); (2) juniors taking the SAT for the first time in April 1976 (247 coached, 617 uncoached); (3) seniors taking the SAT for the second time in November 1975 (98 coached, 396 uncoached); (4) seniors taking the SAT for the second time in November 1976 (177 coached, 387 uncoached); (5) all high school students taking the SAT for the first time on any test date during the three-year period (417 coached, 1763 uncoached); and, (6) all high school students taking the SAT for the second time during this period (316 coached, 1267 uncoached). Statistical analyses were actually based on smaller samples than these largely because of missing student descriptive data.
Since the coached and uncoached groups might differ from each other in a number of ways possibly relevant to the treatment, the demographic and personal characteristics of the two groups were contrasted. It was found that the coached group was significantly higher than the uncoached group in high school class rank, parental income, most recent English grades, most recent math grades, and number of years of math taken; in addition, the coached group included significantly more nonpublic school students and fewer public school students than the uncoached group. Before multiple regression analyses controlling for these and other background variables were conducted, the possibility of differential coaching impact on good and poor students was first discounted by noting a lack of interaction between PSAT scores and coaching treatment.

The multiple regression analyses, which controlled for PSAT (or first SAT) as well as for the several relevant background variables, yielded negligible effects for students at one school and statistically significant effects for students at the other, where the impact for SAT-Verbal was found to be 30 and 27 points, respectively, for first- and second-time SAT takers over the pooled time periods and 19 and 28 points, respectively, for SAT-Math over the same periods. Since these values represent combined coaching and self-selection effects by virtue of the confounding between pre-existing group differences and the coaching treatment, the FTC report then presented an analysis of potential self-selection bias. In a regression analysis of the pooled treatment and control groups, coached students were found to achieve lower PSAT (or first SAT) scores than were predicted from their background characteristics, whereas uncoached students scored slightly higher than expected given their personal and demographic background. This type of self-selection was characteristic of students attending the apparently effective coaching school but was not generally found at the ineffective school. In an effort to eliminate the self-selection effect of this underachievement on the PSAT (or first SAT), the regression analyses were repeated dropping the PSAT (or first SAT) from the set of covariates. As a consequence, the estimated effects were greatly reduced. For example, the effects for SAT-V and -M for first-time SAT takers over the pooled time periods in the previously effective school were 11.5 and .55, respectively, which are no longer statistically significant; the effects for second-time SAT takers over the same period were 16.2 and 16.6, respectively, for SAT-V and -M, which remain statistically significant. It was then argued that this reanalysis would be appropriate.
only if the underachievement on the PSAT for coached students were due to chance. Since an analysis of PSAT scores for students coached between their first and second SAT exams revealed a similar pattern of underachievement on both the PSAT and the first SAT, it was concluded that the phenomenon was not random but was characteristic of students self-selecting coaching, that is, they were underachievers on standardized tests and would likely continue to be test underachievers in the absence of coaching. Therefore, it was argued, the prior results showing coaching effects in the 20- to 30-point range for both SAT-V and -M at one coaching school were the most defensible findings. Those findings still represent estimates of combined coaching and self-selection effects, however, since this attempt to analyze self-selection as underachievement on standardized tests did not alter the confounding of coaching with pre-existing group differences or otherwise eliminate the effects of unmeasured self-selection factors.

The most fundamental issue concerning the FTC coaching study is that it undertakes an interpretation of available data as a substitute for collecting experimental data. The fact that the data are taken from records and files necessarily puts the study out of reach of the kinds of experimental controls that would permit clear, unambiguous interpretation of findings—i.e., it is a quasi-experiment rather than a true randomized experiment. Such a study does not involve random assignment of students to coaching and noncoaching conditions, and in the absence of randomization some interpretive equivocality is inevitable. We hope to reduce—but cannot eliminate—this equivocality by conducting multiple alternative statistical analyses. Summaries of two such reanalyses are included in Sections IV and V of this report and the full texts are reproduced in the Appendix.

**The Power of Randomized Experiments**

The value of a randomized experiment warrants discussion here. To conduct a true experimental study of coaching, one assembles a large group of students representative of the kinds of individuals about whom inferences and generalizations are to be drawn. These students are assigned at random to either a treatment [or experimental] subgroup to receive coaching or to a control subgroup for whom the coaching experience is to be delayed. To increase precision, before the treatment subgroup takes the coaching course, a form of the SAT could be administered to both
subgroups as a pretest. Or, to avoid possible pretest-treatment interactions, a different instrument might be used as a pretest or some other proxy measure of ability or achievement used as a covariate. Effective control conditions should be established to maintain motivation, avoid attrition, and otherwise assure that the two groups remain comparable except that one receives coaching and the other does not. At the end of the coaching period, the SAT is administered as a posttest to both treatment and control subgroups. The experimental data may be analyzed in any one or a combination of several ways. For example, the analysis-of-covariance model uses pretest and other pretreatment variables to adjust for any differences that might exist between the two randomly assigned subgroups on those variables. Since by randomized design and the maintenance of effective controls the only systematic difference between the two subgroups is that one received coaching and the other did not, differences observed on the outcome or posttest measure can be confidently attributed to the coaching experience within some range of standard error.

In a nonrandomized design or quasi-experiment such as the FTC study, in which the coached students were those who had enrolled in coaching schools and the "control" students were drawn from another source, there is no way of discounting alternative reasons for the difference observed on the outcome or posttest measure. The difference might result from the coaching experience or, on the other hand, might simply reflect differences in the characteristics of the two groups existing prior to the coaching experience. One powerful feature of the randomized experiment in this regard is that we can attach probabilities to the likelihood of these alternative events. Thus, although we have other reservations regarding the analysis and interpretation of the FTC study, our chief reservations are those that relate in one way or another to the nonrandomized nature of the study.

**Confronting Treatment and Control Group Differences**

As a consequence of the nonrandom assignment of subjects in the FTC study—indeed, the highly self-selected nature of the treatment group—the data offer numerous opportunities for alternative interpretations. The FTC-BEP (1979) report recognizes the limitations of quasi-experimental data and takes care to avoid the type of simplistic analyses undertaken by the Boston Regional
Office (1978). The BCP analysis takes proper note of the fact that the coached and uncoached groups differ at the one percent level of significance on several key demographic variables relevant to SAT performance, e.g., class rank, parental income, ethnic background, high school type (public or nonpublic), most recent English grade, most recent math grade, and expected years of math.

It is noted in this regard that the BCP analysts took account of certain demographic items available to them from the Student Descriptive Questionnaire (SDQ), but excluded others that seem at least as relevant as those that were included. For example, the SDQ data provided by the College Board and ETS for the study included such variables as students' level of educational aspirations and their parents' level of education, but these variables were omitted on the master tape released by BCP for reanalysis.

There are, in addition, some refinements that might have been introduced in the BCP analysis, but were not:

- It might have proven useful to conduct additional analyses of interactive effects—e.g., the coaching effect with sex of student, with minority status, with the number of previous PSATS taken, the number of years of math study, and high school grades. Analyses to assess possible interactions with treatment have been undertaken by ETS.

- In comparing coached and uncoached students, it would have been informative to report the demographic profiles for each coaching school separately rather than combining them across schools and risking the chance that important school effects might be obscured.

- Although the primary objective of adjusting for the independent variables was fulfilled, it might have been more informative if these variables had been entered into the analysis seriatim rather than simultaneously. This would have made it possible to examine the incremental effects of each variable as it entered into the regression system.

The Bane of Self-selection

We have some major questions regarding the formulation of some of the hypotheses and inferences drawn in the FTC report about the nature and impact of self-selection:
• The FTC report suggests that the control group "may in fact have received some form of coaching other than formal enrollment in a commercial coaching course" [p. 2], in which case the apparent benefit from coaching, as observed in this study, would have been an underestimate of the real coaching effect. However, data from a study by Powers and Alderman [1979] on the effects of using the College Board booklet Taking the SAT suggest that coached students may be more likely than uncoached students to engage in additional methods of preparation. Thus, under these circumstances the FTC analyses would be more likely to overestimate the effects of commercial coaching schools by failing to account for other ways in which coached students may have prepared for the test.

• The hypothesis is offered in the FTC report that students who are coached before taking the SAT for the first time are serious about their test experience and do not plan to take it a second time, while those who are uncoached are not as serious since they do plan to take the test a second time. If this is so, then higher scores obtained by coached students may not be due to coaching but to the difference in how serious the students are. The results showed that coached students were, in fact, not less likely but more likely than uncoached students to take the SAT a second time and the conclusion is drawn that the issue of seriousness is, therefore, not a cause for concern. However, consideration should be given to the counter-hypothesis that because coached students were more likely than uncoached students to take the SAT a second time, coached students are indeed more serious and more determined to persist until they obtain satisfactory SAT scores.

• In considering the possibility of self-selection, PSAT (or SAT) scores were predicted from a composite of demographic characteristics including family income, most recent English and mathematics grades, and class rank, among others, for the pooled sample of coached and uncoached students. These predicted scores were then compared with actual initial PSAT (or SAT) scores. The coached students were found to earn lower initial PSAT (or SAT) scores than expected from their background characteristics, whereas uncoached students, on balance, scored slightly higher than expected given their personal and demographic characteristics. Thus, the coached students on the average achieved lower initial PSAT (or SAT) scores than uncoached students having the same values of demographic varia-
bles. It is concluded that coached students "tend to underachieve on standardized tests" and, further, that coaching is "effective for underachievers." The terminology is rather unfortunate in that it is easy to slip, as happens occasionally in the FTC report, and refer to the coached students as "underachievers"—a term that is generally applied to students whose grades in school are lower than one would expect on the basis of test scores or prior achievement. In fact, the comparison of demographic profiles presented for coached and uncoached students indicates that the coached group earned higher grades in English and mathematics and had higher class ranks than did the uncoached students—but they did not score as well on the PSAT as those accomplishments would have led one to expect.

In contrast to the hypothesis offered in the FTC report that "the students who choose to go to a coaching school represent those individuals who scored lower than they expected on the SAT" (p. 23), a number of other hypotheses about self-selection are at least as plausible and consistent with the data. These include self-selection on the basis of family income, attending private schools, etc. The point is that we simply cannot tell with any certainty on what basis self-selection occurred. For example, students at Coaching School A, which charged $225 for the course, were found to be "underachievers on standardized tests," whereas students at coaching School B, which charged only $75, were not.

- In the FTC regression analyses, the estimated effects of taking the PSAT twice prior to taking the SAT were found to be larger, although less precise, than those attributed to coaching. These effects on SAT performance of repeating the PSAT are unusually high relative to what we know about the impact of prior test practice or test repetition, suggesting the operation of some form of self-selection bias in these data.

- A major conclusion of the study is that "coaching can be effective for those who do not score well on standardized tests" (p. 35). On closer examination, we note that this conclusion is not substantiated in data analysis because the FTC regression analyses, by virtue of not investigating interactions with treatment, estimate the same effect of coaching regardless of demographic characteristics of the subjects. Consequently, there is no evidence in the FTC study on this point of differential effectiveness of coaching. Rather, this conclusion is based on conjecture, as follows: One school produced significant coaching
effects and the other did not; students at the effective School A tended to be underachievers on standardized tests while those at the ineffective School B were not. The FRC report then notes that coaching can be effective for underachievers on standardized tests because it was at School A, rather than arguing that underachievement on standardized tests and attendance at School A are confounded and that self-selection affords a plausible rival interpretation for the obtained effects or for part of those effects. It is then speculated that "if only underachievers can be helped, it is possible that coaching at School B would be effective for such students" (p. 36). Furthermore, in School A self-selection in the sense of "underachievement on standardized tests" was not apparent on one of the test administrations (the second SAT-Verbal taken in 1975), and the conclusion is drawn that "coaching at School A can be effective for all students, not just for underachievers" (p. 35). Although qualified in the FRC report, for conclusions of this importance to be drawn with such generality one would expect both the data and the analyses to be more directly supportive.

**Additional Puzzlements**

The FRC report describes some findings that warrant clarification and further research:

- The effects of coaching on the Verbal section of the SAT were found to be as large or larger than the effects on the Math section. This result seems contrary both to intuition—the Math test is much more curriculum-related than is the Verbal test—and to the general trend of results from earlier studies that the Verbal test is less susceptible than is the Math test to special preparation effects.

- The FRC report gives data showing that the effect for students attending Coaching School A was significantly greater on SAT-Verbal in 1976 than in 1975 and greater on SAT-Math for second-time takers than for first-time takers. These variations are potentially important, and an examination of their sources might illuminate the relative effectiveness of coaching practices.

- The analysis revealed that attendance at School A was associated with larger effects on SAT performance than attendance at
School B. Considering the fact that the tuition at School A is three times the tuition at School B, it is not surprising that the self-selection effects discussed in the report are stronger for the School A students than for the School B students, who might thereby be expected to be more similar to the uncoached students. Such a view would put more stress on the School B findings than is given in the FTC report.

At a more specific level, there is some concern that the procedures used for constructing the samples in the FTC analyses may have introduced biases related to the effects under investigation. For example, the FTC noted that testing histories were identified for 1568 of the coaching school students but not for 600 others. This represents an attrition of about 28 percent, a highly significant loss. In addition, all students who did not respond to the Student Descriptive Questionnaire were dropped from the analysis. The study showed a significantly lower rate of response to the items on the SDQ—approximately 64 and 67 percent for the coached and uncoached students, respectively—than is typical of the College Board candidate group. The usual response rate for the candidate group is about 86 percent.

**New Perspectives on the FTC Data**

In an attempt to clarify some of the points raised in this critique, ETS undertook reanalyses of the data included on the master tape released to the public by the FTC. As part of our effort to understand the conduct and implications of the FTC study, ETS invited Dr. T.W.F. Stroud, Queens University, Kingston, Ontario, to design and carry out additional analyses of the FTC master tape data. As the first step in his independent reanalysis, Dr. Stroud successfully replicated the results as reported by the FTC. Having assured the accuracy of the FTC computations using their statistical design, Dr. Stroud undertook a reanalysis of the data for all three coaching schools using a similar analysis-of-covariance model, but differing from the FTC analysis in two important respects: The FTC analysis employed a pooled regression equation across both coached and uncoached students using, along with other variables, PSAT-V in predicting coaching effects on the SAT-Verbal score and PSAT-M in predicting SAT-Math effects. In contrast, Dr. Stroud employed a regression equation defined by the performance of uncoached students only and, in addition, in-
cluded both PSAT-V and -M in predicting coaching effects for each area on the SAT. This analysis is preferable not only because it is more precisely controlled through the inclusion of additional covariates, but because it more appropriately contrasts the performance of coached students with predicted levels based on uncoached students rather than a mixture of the two. This approach results in valid estimates with fewer assumptions and enables examination of interactions in a straightforward manner (Cochran, 1968).

This reanalysis indicates that, given their background characteristics and pretest levels on the PSAT, students enrolled in one of the three coaching schools studied obtained significantly higher SAT scores than did uncoached students by about 20 to 35 points in both Verbal and Math—the same neighborhood as the FTC estimates. The estimated effects of coaching for the other two schools were not statistically significant. These are estimated combined effects due to coaching and self-selection (self-selection in terms of students who seek and complete commercial coaching programs as opposed to those students who do not), since it is not possible to estimate coaching and self-selection effects separately with these data. No interactions were uncovered at Coaching School A for either SAT-V or SAT-M. At Coaching School B, however, statistically significant and independent interactive effects were obtained on SAT-V for race and self-reported parental income. On the average, even though their number was quite small (N = 13), black students at School B exhibited significantly larger coaching/self-selection effects on SAT-V than non-blacks, and students reporting low family income exhibited significantly larger coaching/self-selection Verbal effects than those reporting high family income. A detailed summary of the Stroud reanalysis follows in Section IV, and the complete report appears in Appendix 2.

A second study of the data was undertaken by Donald A. Rock, Senior Research Psychologist at ETS. Since we know that gains in SAT scores can be expected from the junior year to the senior year of high school, Dr. Rock applied a statistical model incorporating growth effects. For the treatment or experimental group, this study included only those students at the largest coaching school (School A) for whom three sets of test scores were available, a PSAT and two administrations of the SAT. The control group included only those uncoached students for whom these same three sets of test scores were also available. The coached students, who were labeled "underachievers" in the FTC report, performed bet-
ter than the uncoached students on the PSAT and were also higher in high school rank-in-class and family income than were the uncoached students. As would be expected, the coached students also scored higher than their uncoached cohorts on the first SAT administration, but in the case of the Verbal area scored differentially higher. That is, during the period from taking the PSAT to taking the initial SAT, prior to attendance at the coaching school, the verbal skills of the coached students appeared to grow more rapidly than those of the uncoached students. In Math, however, both the coached and uncoached students showed similar group growth rates during this preintervention period. When the confounding effects of differential group growth rates are controlled for, the estimated coaching effect on the Verbal score (about 17 points) is substantially smaller than that for the Math score (about 30 points). The fact that these two effects are different from each other, and not the same, is consistent with the results of earlier studies and with expectations that Math, being generally more curriculum-related than Verbal, might be more responsive to special preparation. A detailed summary of the Rock study follows in Section V, and the complete report appears in Appendix 3.

Most of the criticism in this section points up the limitations of the FTC data and the limits on inferences that can be drawn from such data. Within these limits, further analysis and interpretation may still prove to be illuminating. In this spirit, we next attempt to estimate the combined coaching/self-selection effects using more refined methods and to investigate the possibility of interactions between size of effects and student background characteristics. We also attempt to adjust statistically for that portion of self-selection effects that is embodied in differential group growth rates.
IV

Estimation of Combined Coaching / Self-Selection Effects in the FTC Study: Detailed Summary and Elaboration of the Stroud Reanalyses

In an effort to obtain more precise estimates of the combined coaching/self-selection effects in the FTC data, Dr. T.W.F. Stroud of Queens University, Kingston, Ontario, at the request of ETS, designed and conducted additional analyses introducing several refinements over the procedures used by the FTC. It will be recalled that the FTC study was based on six subsamples: (1) high school juniors taking the SAT for the first time in April 1975 or (2) in April 1976; (3) high school seniors taking the SAT for the second time in November 1975 or (4) in November 1976; and, (5) all high school students taking the SAT for the first time on any test date over the three-year period 1974-1977 or (6) taking the SAT for the second time on any test date during that period. Data for students at two coaching schools in the metropolitan New York area were included, while data from a third coaching school were left unanalyzed because the number of students was considered too small. A control sample of uncoached students consisted of every 150th individual in the ETS files who took the SAT during the given three-year period in the same greater New York area.

The following background variables, roughly in decreasing order of their importance, were controlled for in the FTC regression analyses: pretest score [PSAT-V when the first SAT-V (SAT1-V) was being predicted, SAT1-V when the second SAT-V (SAT2-V) was being predicted, and similarly PSAT-M when predicting SAT1-M and SAT1-M when predicting SAT2-M], self-reported grade in English or Math, self-reported rank-in-class, self-reported years of English or math taken, self-reported parental income, sex, race, high school type (public or nonpublic), and number of PSATS taken. Time between pretest and posttest was also included as a covariate, but it did not improve prediction significantly. The FTC regression analyses were based only on students with complete data on all of these variables. For each of the six subsamples, separate regression equations predicting SAT-V and SAT-M were com-
puted for pooled coached and noncoached students, using dummy variables to represent attendance at Coaching School A and Coaching School B.

Overall, the FTC analyses showed that students at Coaching School A, on the average, scored significantly higher on the SAT than noncoached students. The amount of advantage on a 200- to 800-point score scale falls somewhere between 14 and 38 points, which are 95% confidence limits for the differences in adjusted means (the median lower confidence limit and the median upper confidence limit over the 12 analyses, for SAT-V and -M in the 6 subsamples). Students at Coaching School B did not do significantly better than noncoached students; their score effect falls somewhere between median confidence limits of -12 and +19 points. In any event, since the coached students in this nonrandomized study differed significantly from the noncoached students in a number of ways that could have influenced both their decision to attend coaching school and their performance on the SAT [such as having higher rank-in-class and higher parental income], these obtained score differences between coached and uncoached groups represent a confounding of coaching effects and personal factors that cannot be disentangled with the available data. Among these personal factors influencing attendance at commercial coaching schools, for example, are motivation to earn a higher test score and financial means. The FTC data set includes a rough proxy for financial means in the form of self-reported parental income, but it includes no proxy for motivation or for a host of other important ways in which coached and noncoached students are likely, on the average, to differ, such as in career aspirations or in level of parental education. Attempts were made in the regression analyses to control for student differences in reported income, but there is no way that statistical adjustments can take unmeasured influences into account. As a consequence, the score effects reported in the FTC study must be interpreted as combined coaching/self-selection effects, as must the findings of the reanalyses that follow.

The reanalyses undertaken here differ from the FTC analyses in a number of respects:

In the FTC analyses, verbal pretests (PSAT-V and SAT1-V) and verbal background variables [e.g., grades in English, years of English] were used to predict SAT-Verbal scores, and quantitative pretests and background variables were used to predict SAT-Math scores. However, since the inclusion of both verbal and quantitative variables improved prediction of each score, both verbal and quantita-
tive pretests and background variables were used in the current reanalyses to predict both SAT-1-V and -M for juniors and both SAT-2-V and -M for seniors. The only exception was "number of years of English," which was dropped from the analyses because it did not add to prediction in any regression equation when the other variables were already entered.

The current reanalyses dealt only with the peak test dates in 1975 and 1976 [subsamples 1, 2, 3, and 4 in the FTC study]. Sub-samples 5 and 6 for the pooled time periods were omitted because of their heterogeneity.

Only students with complete data were included in the FTC analyses. In contrast, Stroud used missing-value techniques so that students not reporting parental income, race, or rank-in-class could nevertheless be meaningfully included in the analyses.

In the FTC study, students enrolled in coaching school who did not receive coaching prior to the SAT administration in question were added to the uncoached group. Since the representativeness of the control sample is thereby eroded, these students were excluded from the present analyses.

All three coaching schools in the FTC data set were included in the present analyses, and effects and standard errors are estimated for each school separately. In addition, smoothed estimates are provided for the three schools which utilize the empirical Bayes concept of "borrowing strength" across samples and which allow, under certain assumptions, for the possibility of predicting coaching/self-selection effects in the same schools in future years.

In a key departure from the FTC methodology, a multiple regression equation predicting each dependent variable (SAT-1-V and -M for juniors and SAT-2-V and -M for seniors) is constructed in the present approach on the respective junior and senior samples of uncoached students rather than on a pooled sample of coached and uncoached students as in the FTC study [Belson, 1956]. This procedure yields an unbiased estimate of the treatment effect in the presence of interactions between the size of effects and values of a covariate; it is also the recommended procedure when the control sample is much larger than the treatment sample [provided that the relative contribution of sampling errors in the control-group regression line to the variance of the effect estimate is negligible] [Cochran, 1968]. These regression equations are then applied to the coached students to predict the SAT scores they would have expected had they been uncoached students with their same values on predictor variables. Since we wish to assess
the average score increase of those *coached*, we adjust the *control* group from one mean to the other. Residuals from these regressions are then calculated for students at each coaching school, and the average value of these residuals is taken to represent the combined coaching/self-selection effect for each school. Combined coaching/self-selection effects, together with standard errors (including the sampling errors of the regression line), are reported in Table IV-1 separately for juniors and seniors in each of the two administration years for each of the three coaching schools. The overall results are similar to the FTC findings, although more precisely estimated with smaller standard errors. Combined coaching/self-selection effects were statistically significant for one coaching school and negligible and inconsistent for the other two coaching schools. The significant effects ranged from about 27 to 34 points for SAT-V and from about 20 to 33 points for SAT-M. Smoothed estimates based upon empirical Bayes procedures present a similar pattern in Table IV-2, with the size of the effects somewhat moderated and the standard errors somewhat increased.

**Table IV-1**

Coaching/Self-Selection Effects and Their Standard Errors

<table>
<thead>
<tr>
<th>SAT-VERBAL</th>
<th>1975 SAT1-V</th>
<th>1976 SAT1-V</th>
<th>SAT-MATH</th>
<th>1976 SAT1-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
<td>Effect</td>
<td>S.E.</td>
<td>School</td>
<td>Effect</td>
</tr>
<tr>
<td>Juniors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>27.44</td>
<td>6.66</td>
<td>A</td>
<td>22.36</td>
</tr>
<tr>
<td>B</td>
<td>—</td>
<td>—</td>
<td>B</td>
<td>—</td>
</tr>
<tr>
<td>C</td>
<td>17.88</td>
<td>11.19</td>
<td>C</td>
<td>27.68</td>
</tr>
<tr>
<td></td>
<td>25.64</td>
<td>14.10</td>
<td></td>
<td>7.14</td>
</tr>
<tr>
<td>Seniors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>32.07</td>
<td>5.86</td>
<td>A</td>
<td>22.54</td>
</tr>
<tr>
<td>B</td>
<td>11.12</td>
<td>10.16</td>
<td>B</td>
<td>26.40</td>
</tr>
<tr>
<td>C</td>
<td>-10.47</td>
<td>14.02</td>
<td>C</td>
<td>24.34</td>
</tr>
<tr>
<td></td>
<td>20.30</td>
<td>18.70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another advantage of the present methodology, in addition to yielding valid estimates with fewer assumptions and usually greater precision than the FTC approach, is the straightforward manner in which potential interactions may be explored. This is accomplished in general by correlating the coached students' residual scores, which represent individual coaching/self-selection effects, with whatever student characteristic or background variable one suspects might be associated with differential effects. Variables yielding significant correlations then serve to identify
Table IV-2
Smoothed Estimates of the Mean Coaching/Self-Selection Effects at the Three Coaching Schools

<table>
<thead>
<tr>
<th>School</th>
<th>Effect</th>
<th>S.E.</th>
<th>School</th>
<th>Effect</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.5</td>
<td>9.2</td>
<td>A</td>
<td>16.9</td>
<td>18.6</td>
</tr>
<tr>
<td>B</td>
<td>9.7</td>
<td>12.8</td>
<td>B</td>
<td>7.8</td>
<td>23.0</td>
</tr>
<tr>
<td>C</td>
<td>20.9</td>
<td>10.7</td>
<td>C</td>
<td>8.4</td>
<td>19.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School</th>
<th>Effect</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.4</td>
<td>12.1</td>
</tr>
<tr>
<td>B</td>
<td>9.4</td>
<td>12.3</td>
</tr>
<tr>
<td>C</td>
<td>6.1</td>
<td>15.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School</th>
<th>Effect</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24.0</td>
<td>15.1</td>
</tr>
<tr>
<td>B</td>
<td>17.2</td>
<td>15.3</td>
</tr>
<tr>
<td>C</td>
<td>12.0</td>
<td>17.6</td>
</tr>
</tbody>
</table>

subsets of the coached student population that systematically exhibit greater coaching/self-selection effects than others. As previously suggested, it would not be surprising to find a number of interactions in the FTC data for such variables as "years of math taken," which revealed differential score effects for Dyer (1953)—possibly moderated by sex as it was for French (1955)—or educational or cultural disadvantage as anticipated but not confirmed by Roberts and Oppenheim (1966).

In the present exploratory analysis of interactions, as an overall test, two multiple correlations were computed for the coached students, one predicting the individual Verbal coaching/self-selection effects and the other the individual Math effects, using as predictors the same set of variables that had served as control measures in the initial regression equations computed on uncoached students. In the uncoached group, of course, these variables would be uncorrelated with residual scores. These predictors included PSAT-V and -M (or SAT1-V and -M), high school type, rank-in-class, years of mathematics, self-reported parental income, grades in English and math, sex, and race or minority group membership. (Two minority grouping variables were included: Group 1 refers to Asian Americans, Hispanics, and other nonblack minorities; Group 2 refers to blacks). Only persons with complete data from the three largest cells of coached students were included in this analysis, namely the 1976 juniors and seniors of School A and the 1976 juniors of School B; the sample sizes were 103, 102, and 85, respectively. For the two School A samples, neither the multiple correlation for the Verbal effects nor the multiple correlation for the Math effects was statistically significant. For the School B juniors, the multiple correlation for the Math effects was also nonsignificant but that for the Verbal
effects was significant at the .05 level. An examination of the individual correlation coefficients between each predictor and the Verbal residual score revealed that Group 2 (black vs. nonblack) was the most significant predictor variable. The most important predictor after Group 2 was entered into the equation was self-reported parental income. Group 2 correlated .288 with the Verbal residual score and Income correlated -.250, both of which are statistically significant at the .05 level. The partial correlation coefficient between Income and Verbal residual score after partialling out Group 2 was -.258, indicating that the effect of income and the effect of black vs. nonblack were essentially separate from each other in this sample of School B juniors.

The interaction with Income indicates that among coached juniors at School B those students reporting lower parental incomes exhibited larger coaching/self-selection effects on sat-Verbal than did students reporting higher parental incomes. Since students with low family income may have had less access to learning resources at either home or school, this interaction might have arisen because some of them chose to attend coaching school as a limited-cost effort to gain short-term intensive access to such resources, which they then used to good compensatory effect. Or, as we have previously suggested, since both motivation and financial means are probably instrumental in deciding whether or not to attend a commercial coaching school, it is not unreasonable that coaching school enrollees having lower than average financial means might have especially strong motivation to perform well. It should be noted, however, that no interaction with parental income was uncovered for either juniors or seniors at School A, where the required financial investment is three times that of School B. The reason for such apparent differences in the effects of parental income in School B as opposed to School A is not obvious, unless the nonreplications signify that this interaction is one of those rare random events that achieves statistical significance by chance. On the other hand, the range of income was so restricted at School A—50 percent of the students reported parental incomes in the top category of more than $30,000 a year—that a significant correlation with residual scores would be difficult to obtain.

Plausible interpretations for the interaction with black vs. non-black are even less obvious. For example, at first glance one might expect this interaction to be associated with disadvantage. But if so, it is unrelated to economic disadvantage because the interactive effects of black vs. nonblack in School B
are separate from the effects of self-reported parental income; in this sample, unlike the general population, these two variables are uncorrelated. Nor is educational disadvantage or developed ability level implicated in any simple way, because there were no differential effects associated with pretest score level or with high school grades. Neither was this Group 2 interaction replicated in School A, but in this case it had no opportunity to occur there because only three black students with the requisite background scores were identified in the School A data set.

In an effort to clarify the nature of this Group 2 interaction, test score and background data were examined for all of the coached students who identified themselves as black and who took the SAT either as juniors or seniors during the peak month of either administration year and for whom the information necessary for the missing-value regression equations was available. Fifteen such students were identified, 13 of whom attended Coaching School B. The average Verbal coaching/self-selection effect for these 13 students was 46.7 points above that for white students at School B, which is statistically significant at the .001 level. The average Math effect for these 13 black students also exceeded that for whites, but the difference was not statistically significant. The Verbal effects for the other two black students exceeded the cell mean for their respective coaching school by 24.6 points on the average. Six additional black students were identified who had taken the SAT at nonpeak administrations; their Verbal effects exceeded their corresponding cell means by 52.6 points on the average.

In terms of background characteristics, the total group of black students identified reported parental incomes ranging from 5 thousand to more than 30 thousand dollars per year, with a mean of $18,500. These students, in the main, were in the top third of their high school class and earned grades of B or better in both math and English, but their scores on the PSAT or SAT 1 were relatively low. This pattern of lower test scores than might have been expected from background characteristics is reminiscent of the type of self-selection attributed to School A students in the FTC report, but such "underachievement on standardized tests," it will be recalled, was not found to be typical of School B students on the average. Since the Group 2 interaction might thus be another example of self-selection effects, data were examined for coached and uncoached black students in the Alderman and Powers [1979] randomized study of high school special preparation programs. Only 8 coached black students were identified.
but their individual SAT-Verbal coaching effects were estimated anyway, especially since the interaction in question was based on only 13 black students. Using the same within-school regression equations employed by Alderman and Powers, the black students’ coaching effects were found to exceed their respective school mean coaching effect by only 8.4 points on the average, indicating little differential effectiveness for black students. Using within-school regression equations based only on uncoached white students, as a parallel to the regressions Stroud used for uncoached students in general, the coached black students scored an average of 3.8 points below the level predicted had they been uncoached whites. In contrast, the coached white students scored an average of 9.2 points above what would be expected had they been uncoached. Eight uncoached black students scored an average of 18.2 points below their expectation had they been uncoached whites with their same values on the predictor variables (in this case PSAT-V, PSAT-M, and Test of Standard Written English). The average residual score for the coached black students was thus 14 points higher (−3.8 vs. −18.2) than that of the uncoached blacks, while the average residual for the coached white students was 9 points higher than the expected value for uncoached whites. Hence there is little indication that coaching, at least in the form of high school special preparation programs as opposed to commercial coaching, works differentially better for black students in the absence of self-selection effects.

Whatever the basis for this black vs. nonblack interaction, it does not appear to be minority status in general, because Group 1 (Asian Americans, Hispanics, and other non-black minorities) did not yield a significant interaction. Indeed, when data for some 40 nonblack minority students were examined, their individual coaching/self-selection effects bore more resemblance to those of white students than black students, with the Verbal effect being only about 8 points higher than the corresponding cell means on the average. Tending to elevate that average, however, were 8 Asian American students, 4 at School A and 4 at School B, who exhibited Verbal effects about 46 points greater than their cell means on the average.

One possibility is that the Verbal SAT entails additional dimensions of difficulty for blacks, and possibly for Asian Americans as well, that might be overcome by special preparation—similar perhaps to the test sophistication factor found by Vernon (1962) in the multiple-choice responses of British students relatively unfamiliar with such tests, but not in the responses of American stu-
dents. [This would be particularly plausible if some of the black students in the FTC sample were nonindigenous to the American educational system—from the West Indies, for example.] Counter-
ing this hypothesis, however, are the findings of Rock and Werts (1979), who compared factor structures for sat-V and -M across Native American, Mexican American, Asian American, Puerto Rican, black, and white samples and evaluated the degree to which the same factors were being measured in each population. They concluded that both sat-V and -M were measuring the same things across these populations, in the same units of measurement, with equal accuracy as indicated by equivalent standard errors of measurement. Moreover, on the average, sat-V correlates equally well with college performance for whites and blacks, as does sat-M (Breland, 1979).

Nonetheless, differences in test-taking strategies and in test sophistica-
tion more generally might still be at the root of the Group 2 interaction at Coaching School B, but it is difficult to limit that interpretation to black vs. nonblack differences per se. For exam-
ple, in low-income, inner-city samples that were 75 percent black, Flaugher and Pike (1970) found more random responding on the PSAT than appeared in national administration samples as well as a tendency for the more able students to omit more items, which is counter to the usual trend. When patterns of guessing and omissions on the Graduate Record Examination were con-
trasted for black and white samples matched on total score level, however, no differences in the response patterns of black and white students were obtained (Pike, 1980), suggesting that the previously observed patterns might possibly be characteristic of low-scoring students in general rather than being specific to inner-city blacks.

In sum, the meaning of the Group 2 interaction for sat-V at Coaching School B is both intriguing and perplexing. Even though it is based on only 13 students, the size of the effect was quite large and highly significant statistically. It is clear that well-con-
trolled randomized studies of the possible differential effectiveness of coaching or special preparation for the SAT are needed to clarify this issue and to draw implications for practice.
V

Disentangling Coaching Effects and Differential Growth: Detailed Summary of the Rock Analyses

To recapitulate briefly, the FTC study is a pretest-posttest design in which subjects could not be randomly assigned to experimental and control groups. In the language of Campbell and Stanley (1963), it is a special case of a quasi-experimental design known as the nonequivalent control-group design. In the true experiment with random assignment and effective control, the only rival explanation to the treatment being the cause of the results is sampling error. However, in the nonequivalent control-group design, both sampling error and possible selection bias are plausible rival hypotheses to account for the treatment effects obtained.

The FTC analysis and the Stroud reanalysis both relied heavily on the ability of models in the general tradition of analysis-of-covariance (ANCOVA) to rule out rival hypotheses by making linear adjustments in treatment and control group means for the effects of covarying factors. However, if growth is occurring in the dependent variable (that is, in the verbal and mathematical abilities measured by the SAT), then one of the key ways in which nonrandomized treatment and control groups might differ prior to any coaching intervention is in the rate of this intellectual growth. If the treatment and control groups do differ in growth rates prior to coaching and if these differential rates can be estimated from time series data, even if only linearly from two testing occasions prior to coaching, then another kind of adjustment model can be employed to correct for those self-selection effects embodied in differential group growth that were not predicted from the available covariates used (Bryk & Weisberg, 1976, 1977). Using such a growth model, the reanalysis presented in this section shows in some detail that the Verbal data from the FTC study indeed display differential group growth rates and that estimates of the coaching treatment effect are subject to growth adjustments beyond those predicted from available ANCOVA covariates. The Math data, on the other hand, do not exhibit differential group growth and since the growth model does not adjust for group differences on
background variables unrelated to differential growth, the Math coaching effect is better analyzed by ANCOVA.

More specifically, the FTC analysts did not investigate the possibility of differential group growth rates for coached and uncoached students as a rival hypothesis for part of the score increases observed. Campbell (1969) suggests that if two groups start out at time 1 with divergent means, those with the higher mean very likely mature or grow at a greater rate than those with the lower mean. Campbell calls this the interaction of selection and maturation. The FTC analysis of the group demographic variables provides evidence that coached students represent a self-selected population different from uncoached students in a number of ways, including having higher mean pretest scores. The different groups are members of different populations living in different environments. The different environments interact with differences in ability to create and maintain different levels of performance and different rates of growth, suggesting that the use of a statistical model incorporating differential group growth rates is called for.

An interrupted time series design comprising three data points was used to estimate the relative growth rates of coached and uncoached groups during a period prior to the formal coaching intervention. That is, the following design used observations \(O\) of test performance at time \(t_0\) and time \(t_1\) to estimate growth rate in the absence of a formal intervention \(I\) for both the treatment group \(P\) and control group \(C\). Fortunately, the majority of individuals who attended the most successful coaching school (School A) had three testing sessions, with a coaching intervention taking place between time \(t_1\) and \(t_2\).

<table>
<thead>
<tr>
<th>Coaching Program Group (P)</th>
<th>(O_0)</th>
<th>(O_1)</th>
<th>(I)</th>
<th>(O_2)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Control (C)</th>
<th>(O_0)</th>
<th>(O_1)</th>
<th>(O_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t_0)</td>
<td>(t_1)</td>
<td>(t_2)</td>
<td></td>
</tr>
</tbody>
</table>

Figure V-1 presents means and standard deviations of the Verbal test scores at three points in time. Similar data are presented for Math test scores in Figure V-2. The three testings consisted of the PSAT, the SAT taken for the first time in April as a junior, and the SAT taken for the second time in November as a senior. Only one of the three coaching schools in the FTC data set had a large enough student population to furnish sufficient subjects with all three data points. The coached sample included 192 individuals while the control sample numbered 684.
The pattern of verbal data in Figure V-1 suggests that the observed group differences at posttest may well be reflecting differential growth rates as well as treatment effects. Not only is the classic growth phenomenon of increasing means demonstrated in Figure V-1 for both the coaching program students and the controls, but it can be seen that the slope of the line from the PSAT administration \( t_0 \) to the first SAT administration \( t_1 \) is considerably steeper for the coached group than for the control group. The difference in mean score from \( t_0 \) to \( t_1 \) for the coached group is 22 points while the difference for the uncoached control group is 10.
points—a discrepancy of 12 points. On the other hand, while there appear to be significant differential group growth rates in the Verbal area, this does not seem to be the case for Math (see Figure V-2). This is not unreasonable since mathematics is a skill which requires practice if students intend to maintain or increase their level of achievement. In sat-Math, the slopes of the lines for coached and control groups are nearly parallel, with the mean difference in Math score from $t_0$ to $t_1$ for the coached group only 3 points larger than that for the control group.

In addition, mean scores for the national College Board sat
sample tested in April of the junior year \(t_1\) and retested in November \(t_2\) of the senior year are also plotted in Figures V-1 and V-2. These plots represent the 1975 and 1976 populations. Inspection of the verbal plots (Figure V-1) suggests that gains of 12 to 17 points from April to November appear to be commonplace for junior to senior retesters. The fact that the control sample shows a gain of 11 points over the same interval from \(t_1\) to \(t_2\) suggests that they are reasonably representative of junior to senior retesters. Similarly, in Figure V-2, the April to November gain in sat-Math for the control group is consistent with the 17-point gain for the 1975 and 1976 junior to senior retesters in the national sample. The control group’s rate of gain is consistent and linear over both plotted time periods \(t_0\) to \(t_1\) and \(t_1\) to \(t_2\).

An adjustment index may be derived from these data which takes into consideration differential group growth rates in the absence of formal intervention. If the coached group grows at a faster rate than the control group prior to intervention, as might be expected by virtue of its higher initial mean score, the adjustment index \((b^*)\) in the growth model is greater than unity, in contrast to the adjustment index \((b)\) in the traditional ANCOVA model for test-retest data which is typically somewhat less than unity. Using adjustment indices derived from the FTC data, treatment effects were estimated for the growth model and contrasted with the effects obtained from the standard ANCOVA model.

Table V-1 presents ANCOVA and growth-model adjustment indices and coaching effects, as well as the mean growth rates for the coached and control students for Verbal and Math. These estimates of growth are based on the time period between the first and second testing \((t_0\) to \(t_1)\) and thus are not confounded with the coaching intervention. The estimates of growth rates assume linearity of growth, which would seem to be a reasonable assumption given the restricted time period and, even more importantly, given that the observed data for the control group on both Verbal and Math conform quite closely to a linear growth model. If one were to expect a deviation from linearity, it would more likely be in the direction of a steeper slope (in the absence of intervention) between the second and third testing since the motivational level would be at least as great during this time period as in the previous time period between taking the PSAT and the junior-year SAT.

The estimates of the coaching effect under the growth model are about 17 points on Verbal and 31 points on Math. This finding that the Verbal effect is substantially less than the Math effect is
in marked contrast to the ANCOVA estimates of quite comparable V and M effects. The growth model appears to yield an estimate of the Verbal coaching effect that is more consistent with earlier studies and with expectations that Math, being generally more curriculum related than Verbal, might be more responsive to coaching or special preparation.

The results for sat-Verbal in Table V-1 call into question the adequacy of the standard ANCOVA to correct for selection effects when they are present in the form of differential group growth rates not predicted from available covariates. The ANCOVA adjustment index \( b \) is approximately 1.0 for the Verbal data and somewhat paradoxically 1.25 for the Math data, which is the one situation displaying little or no evidence of differential group growth. However, since the growth rates appear to be static for the Math data, the ANCOVA model is probably the more defensible approach for estimating the coaching effect for Math because ANCOVA controls for all available measures of pre-existing difference, not just those related to differential growth.

In summary, then, examination of testings at three points in time suggests that: [1] the traditional ANCOVA approach used by the FTC is inadequate for the Verbal data because of self-selection effects which are at least partially captured in differential group growth rates; [2] a more appropriate growth-related adjustment

<table>
<thead>
<tr>
<th>Table V-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Growth Rates, Adjustment Indices, and Estimated Coaching Effects Under Different Model Assumptions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Average Growth Rate(^1) (Points/Month)</th>
<th>Adjustment Indices and Their Estimates of Coaching Effects(^1)</th>
<th>Growth Model</th>
<th>ANCOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu )</td>
<td>( b ) ( a )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>3.294</td>
<td>1.485 16.946</td>
<td>.993</td>
<td>28.065</td>
</tr>
<tr>
<td>Control</td>
<td>1.654</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>2.354</td>
<td>1.102 30.558</td>
<td>1.252</td>
<td>27.625</td>
</tr>
<tr>
<td>Control</td>
<td>2.048</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)The average gain in points per month is estimated in the absence of intervention. Growth model estimates \( b \) and \( a \) are the adjustment index and the estimated effect based on the group growth rate estimates in the first column. The ANCOVA estimates are the standard estimates arrived at using the list of control variables presented in Table 2, Appendix 3. The \( a \) in the usual ANCOVA equation it yields the effect estimate \( a \).
model yields Verbal coaching effects about one-half the size reported by the FTC; and, [3] the Math data are more consistent with the ANCOVA model. Thus, the resulting FTC estimate of the coaching effect in Math is more likely to be reasonable given the available control variables than is the growth model estimate, since the latter does not adjust for group differences in background variables unrelated to growth. This is not to say that the FTC estimates of the Math coaching effect, as well as those of the present analyses, are not overestimates (or underestimates), since the only self-selection causes that have been adjusted for were those reflected in differential growth rates and/or available demographics.
VI

Implications for Educational and Testing Policy and Practice

In summary, the FRC study of commercial coaching found negligible effects for students attending one coaching school and combined coaching and self-selection effects of about 20 to 30 points on both SAT-V and -M for students at another school. The reanalysis by T.W.F. Stroud, using a more sophisticated analysis-of-covariance design and all three coaching schools in the FRC data set, yielded similar overall results: combined coaching/self-selection effects in the neighborhood of 20 to 35 points for both SAT-V and -M at one school and inconsistent and negligible effects at the other two schools. The effects due to coaching per se at the one apparently effective school are probably lower than this, however, because of the confounding with self-selection factors that influence both attendance at that coaching school and performance on the posttest SAT. As we have seen, for example, when factors related to differential group growth rates in the treatment and control groups were taken into account in the reanalysis conducted by D.A. Rock, the combined coaching/self-selection effect for SAT-V dropped to about 17 points while that for SAT-M, which did not exhibit differential group growth rates in these data, remained at about 30 points. But it is impossible to determine with any confidence whether the effects obtained in the FRC study may be attributable in whole or part to uncontrolled self-selection factors rather than to any impact of the coaching program as such.

Thus, overall, the FRC study appears to reveal considerable variability in the coaching/self-selection effects associated with coaching-school attendance, with an estimated combined effect for students at one school of about 20 to 30 points on SAT-Math and very likely about half to two-thirds that (as reflected in corrections for differential group growth rates) on SAT-Verbal. In addition, the sporadic emergence of significant interactions indicates that particular types of students, such as those highly motivated to achieve, or students with particular cultural backgrounds, such as blacks or Asian Americans, might sometimes exhibit larger score increases in some commercial coaching programs. Further research is needed on this problem since the samples of
blacks and other minorities in the FTC study were exceedingly small and somewhat atypical. On balance, however, the overall findings for commercial coaching appear generally consistent with the results of prior studies on the effectiveness of special preparation programs offered by high schools, especially the longer and more intensive ones.

To pursue this latter conclusion in more detail, let us inquire how the results of the FTC reanalyses jibe with the rankings of prior coaching studies in regard to student contact time and magnitude of score effects, as summarized in Table II-3. The coaching program at School A in the FTC study entailed 40 hours of student contact time while that at School B involved 24 hours. If it is assumed that roughly half that time in each case was devoted to Verbal coaching and half to Math coaching, then the Verbal student contact time of School A would receive a rank of 8 when added to the rank-order correlations of Table II-3, while the associated Verbal combined coaching/self-selection effect (using the weighted average of estimates from Table IV-1) would receive a rank of 7. The rank for the Math student contact time at School A would be 5, while that for the associated Math combined coaching/self-selection effect would be 4. The corresponding values for School B would be rank 10 for Verbal student contact time and rank 13 for combined SAT-V effect, and rank 7.5 for Math student contact time and rank 11 for combined SAT-M effect. The consistency in these associated ranks between student contact time and magnitude of combined coaching/self-selection effects, especially for School A where the associated ranks differed by only one ordinal position, suggests that the FTC results are of a comparable order with the results of prior studies. With the values for School A included, the new correlations between rank-order of student contact time and rank-order of score effect are .77 for SAT-V across 18 studies and .78 for SAT-M across 11 studies. Further, if the statistically unreliable values of School B are added, the new correlations become .77 for SAT-V across 19 studies and .74 for SAT-M across 12 studies.

If it were to be shown that relatively intensive coaching could substantially improve scores for some students on the SAT, this would have important implications for both educational and testing practice. In considering what these policy implications might be, it would be important to know whether any increased test scores attributable to coaching represent stable improvements in the verbal and mathematical reasoning abilities measured by the SAT or whether they reflect improved facility in overcoming inadvertent sources of test difficulty unrelated to these reasoning abil-
Ities, such as difficulty associated with test anxiety or unfamiliarity with different item formats or test-taking strategies. The former case would threaten neither the interpretive validity nor the predictive utility of the SAT, for any genuine long-term improvements in the abilities measured should lead to score increases and should also serve the student well in coping with the demands of college learning, which typically entail the same kinds of verbal and quantitative reasoning by analogy, induction, and deduction as are measured by the SAT. The latter case, by virtue of pointing to construct-irrelevant test difficulty, has double-edged implications: If improved test wiseness increases test scores that were inaccurately low because of construct-irrelevant difficulty, a more accurate assessment of ability level would result and the predictive value of the test should be enhanced. On the other hand, if improved test wiseness leads to test scores that are inaccurately high, the interpretive validity of the test would be diluted and its predictive validity jeopardized. In any particular instance, of course, some combination of improved abilities and improved test wiseness might contribute to improved test performance.

None of the studies of coaching or special preparation conducted so far has systematically addressed this issue of improved abilities; rather, they have all focused on the prior issue of first identifying any score increases associated with coaching programs before inquiring into their causes. This second step has never been taken and, as a consequence, at this point we can only conjecture as to the likelihood that obtained score increases reflect improved reasoning skills. On the one hand, the verbal and

*Although no coaching study has directly addressed this basic question of whether score increases reflect improved abilities or increased test wiseness, Marron (1965) did broach the issue indirectly by investigating the extent to which SAT scores after special instruction predicted freshman class standings at the U.S. service academies and at selective colleges. This appraisal is indirect because it requires not only that the verbal and mathematical reasoning abilities be improved by instruction but that they be effectively utilized in college performance, whereas there are numerous countervailing factors, athletics and extracurricular activities among them, that contribute to scholastic underachievement. Thus, high correlations would provide positive evidence that increased test scores reflect improved abilities being well utilized in college performance, but low correlations would constitute negative evidence only if plausible rival explanations for poor academic performance could be discounted. As a further complication, relatively crude and approximate methods of equating were required in Marron's study to obtain some semblance of distributional comparability across the service academies and across the widely diverse colleges that the preparatory students dispersed to, thereby rendering the outcome admittedly tenuous. Nonetheless, the findings suggest that at the service academies the students did less well academically than the test scores predicted, whereas at the selective colleges the distributions of class standings and test scores did "not seem to be inconsistent." [p. 22].
mathematical reasoning abilities measured by the SAT develop over many years of experience and use in nonschool and school settings, and it is difficult to improve them markedly with short-term interventions in late adolescence or young adulthood. On the other hand, the more productive special preparation programs were relatively intensive—the one commercial coaching school where students exhibited significant average effects, for example, involved 10 four-hour sessions plus homework, which is about the equivalent of a one-semester college course. And the effect on scores, while valued, was not marked. Whether due to coaching or other factors, 20 score points or so on a score scale ranging from 200 to 800 with a standard deviation of 100 amounts to fewer than three additional items correct. It is thus quite consistent with the notion of gradual development of reasoning abilities to argue that intensive study over many sessions might yield modest but genuine improvements in reasoning skills.

It would be important to investigate whether, and to what extent, coaching or intensive special preparation programs may lead to improved abilities and, in any event, to consider how secondary schools might be able to improve their students' academic abilities over the longer haul. Since reasoning abilities facilitate educational accomplishments, school programs stressing ability development as well as subject-matter learning should have a synergistic effect on current and subsequent educational achievement. The verbal and mathematical reasoning abilities measured by the SAT are related, after all, not only to college performance but to high school performance as well. The function of the SAT is to measure validly and uniformly across different groups and different settings the current level of developed scholastic abilities facilitative of academic learning, in an effort to improve the accuracy of forecasting subsequent academic performance at the college level. Whether the current level of developed abilities facilitative of college learning derives in part from intensive special preparation programs pointed toward test performance or from long-term general preparation programs pointed toward school attainment or from extensive experiential learning in nonschool settings—that is, from coaching or instruction or experience—is indistinguishable to the SAT. The issue of equity of access to coaching programs that are effective by virtue of ability development, if such could be identified, is thus similar to the issue of equity of access to effective school programs or effective life experiences. Thus, coaching or special preparation programs producing increased test scores by improving the abilities measured
would have important implications for educational practice and social policy.

In contrast, if coaching or special preparation programs may lead to substantially increased test scores without improving the abilities measured, there are important implications for testing practice. Such an outcome would imply that the test or the testing experience entails unintended sources of difficulty, such as anxiety over being evaluated or unfamiliarity with different item formats or test-taking strategies, that can be overcome by special preparation. Issues of equity of access to such special preparation become important to the extent that individual differences in test-taking skills per se influence test scores.

Although one might expect the typical coaching program to reduce unfamiliarity by drilling the student in different approaches to different item formats and to allay anxiety by providing feedback on effective item performance, the fragmentary research evidence suggests that if this is all that is done the coaching will have little impact on test scores. The coaching or special preparation programs that emphasize drill-and-practice appear to be associated with small effects, if any, whereas programs that include skill-development components tend to be associated with larger effects (which, by and large, seem to occur more for Math than for Verbal). Nonetheless, following accepted principles of good test practice, test makers should strive to reduce construct-irrelevant test difficulty wherever possible—for example, by avoiding arcane item types, complicated instructions, esoteric or culture-specific content, undue speededness, and the like. Substantial efforts should also be made routinely to increase test familiarization and test wiseness, such as the College Board has done by distributing a sample SAT and answer key to all candidates along with strategic advice in the booklet Taking the SAT. With well-made tests the concern is not that students will learn how to capitalize on extraneous clues to inflate their test scores, for such clues should be rare and are further reducible through test analysis, but rather that students learn to cope most advantageously with the test as a standardized vehicle for demonstrating their abilities. To ensure the fairness of the admissions testing process, then, test makers should attempt to make clear to all prospective test takers what the test measures, what kinds of items and formats will be encountered, and what kind and level of preparation the test presupposes, as well as providing practice materials and advice on effective test taking. This should be done not just to minimize the effects on test scores of student
differences in this regard, but to increase the validity of the test and to enhance the value of the testing experience for each individual.

Finally, what advice does all this imply for the prospective SAT taker? If this advice were based mainly on a summary of the research findings as best we can interpret them at this point, the student might be told something like this:

The verbal and mathematical reasoning abilities measured by the SAT develop over years of experience and use both within and outside of school. In preparing for the test, a review of mathematical concepts may be beneficial, especially for students not currently enrolled in mathematics courses.

Studies of the effects of drill-and-practice on sample items or of last-minute cramming show little score increase, although such practice may serve to familiarize the student with different item formats and to reduce apprehension about what to expect. The College Board distributes test-familiarization materials to all candidates and recommends that they take the sample SAT provided, carefully review the answers, and read the booklet on Taking the SAT.

Studies of students attending special preparation programs available in many high schools show a range of effects averaging less than 10 score points on Verbal and less than 15 points on Math for programs requiring an average of about 20 hours or so of total student contact time. Part or all of these score increases may be due to factors other than the special preparation program offered, such as the level of motivation or prior educational experience of the students. And individual students or certain groups or types of students might experience larger increases or possibly even losses.

A recent study of students enrolled in commercial coaching courses shows variability in average score improvement depending on which course the students attended: negligible effects were found for students attending two of the three coaching schools studied, while average effects in the 20- to 30-point range for Math and perhaps half to two-thirds that for Verbal were associated with attendance at the third school, which entailed about 40 hours of total student contact time. Even for students attending this latter school, however, it was not at all clear what proportion of the score increases was due to coaching experiences as such rather than to motivational and other background characteristics typical of the students taking that course. Again, individual students or certain groups or types of students might experience larger increases or possibly even losses.

There have been no systematic studies comparing the relative effectiveness of different types of coaching programs, whether they be commercial or school-based, let alone their relative effectiveness
for particular groups or types of students, but the larger score effects appear generally to be associated with longer-term intensive programs emphasizing skill development in verbal and mathematical reasoning.

In any event, whether due to coaching or to personal factors, score increases of 20 to 30 points on a 200- to 800-point scale correspond to approximately three additional items correct, and comparable or larger effects might be obtained if students were to devote the same level of effort on a regular basis to classroom instruction and independent study.

Although further research is needed to clarify the nature of the relationship, it seems very likely that improvement of the comprehension and reasoning skills measured by the SAT, when it occurs, is a function of the time and effort expended. Furthermore, each additional increase in ability may require increasing amounts of time and effort, probably geometrically increasing amounts. But since the time required to achieve average score increases greater than 20 to 30 points or so, especially in Verbal, may rapidly approach that of full-time schooling, the soundest long-range mode of preparation for the SAT would appear to be a secondary-school program emphasizing the development of thought as well as knowledge.
References


Stern, J. Personal communication, 1975.


VIII
Appendices
Appendix 1

Critical Notes on the FTC Coaching Study

Critical Notes on the FTC Coaching Study

DONALD L. ALDERMAN, November 16, 1979

Staff Memorandum of the Boston Regional Office (BRO)

The disclaimers and reservations attached to the Boston memorandum make it quite clear that the Federal Trade Commission (FTC) itself found serious flaws in this staff report. Yet it may be useful to enumerate those criticisms as well as to offer further comments on specific details since the "revised statistical analyses" conducted by the FTC's Bureau of Consumer Protection (BCP) in Washington accepts the data set and repeats several assumptions of the Boston memorandum.

The note on the cover of the Boston memorandum goes further than usual disclaimers in stating that "the Commission specifically believes that some of the conclusions in the study are not supported by the evidence obtained in the investigation." This strong disavowal of certain conclusions may be attributed to "several major flaws in the data analysis," quoting from the BCP's notice to recipients of the Boston memorandum. The notice cites four specific flaws: 1) comparisons are made "for groups of coached and uncoached students without controlling for differences which may exist in personal and demographic characteristics of the students in the two groups;" 2) failure "to provide tests of statistical significance which are necessary to interpret the results" (i.e., whether differences may simply be due to chance occurrences rather than treatment effects); 3) defects that concern "the method used to present the findings from the data analysis" (e.g., discussion of results in terms of a nonexistent subpopulation of students); and, 4) "all the limitations associated with a nonexperimental design."

These points stressed by the FTC focus on the statistical analysis and give inadequate attention to the data editing which necessarily took place prior to the stage of statistical analysis. Moreover, the subsequent BCP report repeats several weak assumptions made in the Boston memorandum.
Assumptions and Data Analysis

These additional comments concern some of the BRO's assumptions and some of the steps taken in constructing the final data base. But the design of the study would itself preclude any definitive conclusions. Numerous alternative explanations exist for any differences in the test performance of the two comparison groups (i.e., "coached" and "uncoached") since students in these groups differ markedly on key demographic characteristics (see Table 1 of the BCP's report, pp. 8-11). And the statistical analyses undertaken in the BRO memorandum exacerbate these initial group differences by ignoring them. Nevertheless, I offer a few additional comments:

- **Scope of inferences.** The BRO memorandum assumes that "valid inferences about the coachability of other examinations can be drawn from the specific results we obtain for the SAT and LSAT." Given the pre-existing differences evident in the comparison groups, there is not even a strong basis for inferences about the SAT and LSAT let alone any other examinations. Indeed, the results themselves show inconsistencies across test administrations, commercial schools, and examinations.

- **Representative sample.** The enumeration of the study's assumptions [BRO memorandum, p. 49] begins with a statement concerning the sample's representativeness of the entire SAT and LSAT candidate populations. Although the control group was a random sample drawn from history files, the "coached" group is obviously very different from the population of candidates [see Table 1 of BCP's report]. The release of BRO's technical appendix "SOQ," which gives demographic profiles by comparison group, should confirm the weakness of this assumption.

- **Consistency of treatment effects.** Another assumption was that the "coaching school [effect] is consistent during the study period [p. 49]." Certainly the consistency of treatment effects across schools or test administrations should be an open question subject to investigation rather than conjecture or assumption.

- **Treatment self-selection and control contamination.** The strangest assumption made in the BRO memorandum is that "the effects of enrollee self-selection, if any, and of coaching of presumably uncoached students offset one another [p. 49]." There is no way to estimate the extent of treatment contami-
nation in the control group since there were no data collected on this issue, and the BRC memorandum made no attempt to identify a component of score change which might be attributed to self-selection in the treatment group. Yet the statement appears that these unknown and unspecified effects offset one another.

- **Data editing.** The steps which led from the ETS tapes and the commercial schools' enrollment rosters to a final data base were difficult to follow in the BRC memorandum. If my reconstruction of these steps is accurate, substantial numbers of students enrolled in commercial coaching courses were omitted from the final data base:

<table>
<thead>
<tr>
<th>Enrollment in commercial course</th>
<th>LSAT</th>
<th>SAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matches with ETS test files</td>
<td>11,906</td>
<td>2,286</td>
</tr>
<tr>
<td>Students with results from one or two test administrations</td>
<td>9,029</td>
<td>1,777</td>
</tr>
<tr>
<td>Students with GPA (LSAT) or PSAT (SAT)</td>
<td>6,894</td>
<td></td>
</tr>
<tr>
<td>Students with known sex and ethnicity</td>
<td>4,662</td>
<td></td>
</tr>
<tr>
<td>Final &quot;study group&quot; in treatment (coached) condition</td>
<td>2,830*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3,509*</td>
<td>603</td>
</tr>
</tbody>
</table>

*p. 58*

Such sizeable losses (and gains!) raise questions about possible changes in the nature of the sample at each step of the data editing. It has already been noted elsewhere that the number of commercial school enrollees without SAT testing histories seems high (N = 509). But the reduction in sample size from the number of cases with testing histories to the final "study group" is twice as large.

*Revised Statistical Analyses*

*of the Bureau of Consumer Protection*

The BCP report is a much stronger technical document than the BRC memorandum. Its conclusions must be subject to the same qualifications since the study necessarily involves comparisons against a nonequivalent control group, but the BCP report acknowledges these limitations and takes care to qualify any inferences from the results. Furthermore, the explanation of the steps
taken in editing and in analyzing the data is far better than that contained in the BTO memorandum. Nevertheless, they indicate that SAT testing histories were identified for 1568 coaching school enrollees but were not located for "approximately 600 individuals identified from the coaching school list" (p. 5). This attrition rate is about 28 percent and represents a significant loss. Moreover, the response rate for the SDQ for the coached and uncoached students (64 and 67 percent, respectively) is considerably lower than that typically found for SAT candidate groups—about 86 percent. Such high attrition rates from each source may have introduced biases relevant to coaching and its effects.

The BCP report, despite its inherent limitations, provides information on a topic of concern and interest. On the whole, we should welcome the report as a contribution to the literature on coaching. The effects (see BCP report, Table 3, p. 20) correspond to three additional items correct on each of the SAT-V and SAT-M and probably incorporate several components, including self-selection, instruction, and coaching. Moreover, those effects held just for School A. School B had negligible impact on SAT scores. If we view or demonstrate these effects as probable overestimates arising from comparing nonequivalent groups and confounding several possible sources of score changes, three additional items correct seems not to represent a dramatic gain. The same effect might be found for students devoting comparable effort and time to independent study or regular academic courses covering typical subject matter.

Comments on selected points in the BCP report follow:

"Underachievers." The BCP report suggests that coaching at School A was "effective" for "underachievers." An estimated treatment effect of 25 points was found at School A for students whose SAT scores were lower than might be expected on the basis of their high school grades, class rank, and other demographic characteristics. Usually students whose school grades seem high in relation to their test scores would be called "overachievers." And that term, "overachievers," seems more consistent with the demographic profiles in Table 1 (BCP report, pp. 8-11); coached students had higher class ranks (31% in the top tenth of their class), reported larger parental incomes (41% with $30,000 or more), attended nonpublic schools more often (45% in nonpublic schools), had higher English grades (55% with an A in their latest course), and had higher mathematics grades (48% with an A in their latest course) than control students. These students in the coached group were not underachievers or disadvantaged, they were overachievers and middle-class.
**Adjustment for self-selection.** The dropping of PSAT as a covariate (and the concomitant increase in the standard error of the estimated treatment effects) in attempting to compensate for self-selection left me puzzled. After confirming students’ self-selection into the coached group by noting unaccounted differences in PSAT [BCP report, Table 4, p. 26], PSAT is dropped as a covariate in the regression analyses. If we extend that argument to all other independent variables which show significant differences between the coached and uncoached groups [BCP report, Table 1, pp. 8-11], the only covariates left would be sex and expected years of English. The point is that all of the variables, dependent and independent, show self-selection and no statistical analysis will fully compensate for the nonequivalent control group.

**True experimental design.** The BCP report takes the position that a true experiment with random assignment of students to comparison groups would be unethical (“as it would require denying access to commercial coaching to students who want it” [BCP report, p. 3]), expensive, and time-consuming. But our own most recent study of special preparation for the SAT-V did follow a true experimental design by delaying rather than denying access to the treatment.

The FTC memorandum from the BRO and report from the BCP do not close the question of the impact of commercial coaching enterprises. It has not been unequivocally shown that such programs cause score gains on national admissions tests. Indeed, no amount of research can disprove the possible effectiveness of coaching. The possibility will remain open since there is always another program or a different student population. Even an experimental study showing no significant effects this year would not be conclusive evidence regarding new programs and different students next year. But we can perhaps clarify this issue by explaining sources of score changes on major admissions tests and by studying the impact of regular academic programs.
Critical Notes on the FTC Coaching Study

DONALD E. POWERS, November 19, 1979

The claims made for the efficacy of special preparation or "coaching" by those commercial or proprietary schools that offer such preparation programs for standardized admission tests is a proper concern of the Federal Trade Commission (FTC). The Commission deserves support for addressing this thorny issue and for assembling what appears to be the first credible set of data from which inferences about the effects of commercial coaching reasonably could be attempted. As a result of the FTC effort we now know something more about the effects of commercial coaching. Through further analyses that are possible with the FTC data base, even more may be learned.

The FTC should also be applauded for acknowledging the limitations inherent in the type of quasi-experimental study that was conducted, for recognizing the serious flaws in an earlier report issued by its Boston Regional Office, and for attempting to rectify some of these flaws through reanalyses.

Nevertheless, with respect to the more technically adequate report issued by the FTC's Bureau of Consumer Protection (BCP) in Washington, a number of comments can be made regarding: [1] the general reporting of results; [2] the assumptions underlying various analyses and findings; [3] the design and analysis strategies employed; [4] the testing of hypotheses; [5] the logic involved in drawing inferences; and, [6] the reasonableness of results. Needless to say, perhaps, these categories may have substantial overlap.
Reporting

In terms of style, the BCP report is generally well-written, with a minimum of technical jargon. However, four relatively minor points might be made.

- The report's title, "Effects of Coaching on Standardized Admission Examinations," seems to suggest that the results of the study pertain to a variety of standardized admissions tests, which include both aptitude and achievement measures, and to a variety of coaching methods and courses. Since the study has been restricted to specific types of coaching and to one particular aptitude test (the SAT), a title more adequately reflecting the narrower focus of the study would have been more appropriate.

- The report's executive summary is selective in focusing on the results at only one of the two schools involved. Inclusion of a statement concerning the lack of effectiveness of the second school would have been appropriate.

- The report's summary of an average effect of 25 points "on both the verbal and math SAT exams" at one school obscures some important information about the effects, namely that the effect on Verbal scores was somewhat higher than the effect on Math scores.

- Detail for some especially relevant descriptive information is omitted from the report. Table 1 of the BCP report, which shows demographic profiles of coached and uncoached students, fails to provide the statistics for PSAT scores, which are acknowledged in the narrative to differ for coached and noncoached students. The omission of data on the PSAT is glaring because of the report's contention that self-selection is based primarily on these earlier test scores.

Assumptions

Like most quasi-experiments, the FRC study rests on a number of assumptions, both explicit and implicit, that can be examined for reasonableness:

- The BCP report states on page 2 that:

  Conversely, a non-experimental control group may lead to an un-
de-restimate of coaching benefits, that is, the uncoached students may in fact have received some form of coaching other than formal enrollment in a commercial coaching course. They may, for example, have attended a course offered by a school in the not-for-profit segment of the industry or have engaged in extensive self-preparation. These unmonitored efforts, if they occur and if they are effective, would tend to increase the average test scores of the "uncoached" students. These increased scores, containing a component properly attributable to coaching, would tend to shrink the apparent benefit from commercial coaching.

In the absence of any data, this assumption would seem reasonable. However, some reworked data (Table 1) from the study of the effects of the booklet Taking the SAT suggest that coached students may be more likely than uncoached students, except for the use of College Board publications, to also engage in other methods of preparation. Thus, by regressing reasoning, their analyses would actually be more likely to overestimate the effects of coaching by failing to account for the other ways in which coached students may have prepared.

Moreover, the fact that coached students may prepare more diligently than uncoached students would seem to suggest a motivation to do well on the SAT that was probably not captured by any of the variables included in the rrc study. (It should be noted that the coached students in Table 1 reported "attending coaching sessions outside school," but not necessarily commercial coaching schools of the type studied by the rrc.) Hence, the reasonableness of this rrc assumption that preparation efforts on the part of the uncoached students tended to shrink the apparent benefit from commercial coaching may be questioned.

- The assumption that, compared to nonexperimental studies, experimental studies of the effects of special preparation are impractical, unethical, expensive, and time-consuming can be questioned by citing the recently completed experimental study of special preparation programs offered by secondary schools by Alderman and Powers.

- The rrc study attends to the importance of self-selection, but apparently implicitly assumes that all students use the same basis for choosing to attend coaching schools, i.e., that self-selection is uniform. This assumption attributes a simplicity to self-selection that may not be warranted.
Table 1
Use of Various Preparation Methods by Coached and Uncoached Students

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Coached</td>
<td>46</td>
<td>74.2</td>
<td>28</td>
<td>82.3</td>
<td>34</td>
<td>54.8</td>
<td>43</td>
<td>69.4</td>
<td>10</td>
</tr>
<tr>
<td>N = 62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncoached</td>
<td>978</td>
<td>76.5</td>
<td>563</td>
<td>92.0</td>
<td>361</td>
<td>28.2</td>
<td>545</td>
<td>42.6</td>
<td>110</td>
</tr>
<tr>
<td>N = 1279</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1024</td>
<td>76.3</td>
<td>591</td>
<td>91.5</td>
<td>395</td>
<td>29.5</td>
<td>588</td>
<td>43.8</td>
<td>120</td>
</tr>
</tbody>
</table>

\[ x^2 \] 1.69
\[ p \] p = n.s.

Source: Powers, D. E. Students' Use of and Reactions to Alternative Methods of Preparing for the SAT. In progress.

1Base is 34. 2Base is 612. 3Base is 646.
Design and Analysis

The FTC presents the inherent limitations of the non-experimental nature of the design used. The application of multiple regression analysis as the primary data analysis technique seems appropriate. However, the following additional refinements in analysis might have been useful:

- A more careful entry, perhaps serially, of variables or sets of variables in the analyses, instead of entering all of them simultaneously, might have been even more informative. A careful examination of the collinearity of variables might also have been desirable.

- Additional analyses of possible interactive effects might have proven useful. The FTC researchers did test for PSAT coaching interactions and concluded that they need not be taken into account. However, a number of other interaction effects could have been included in the analyses, such as the interaction of coaching with:
  —number of previous PSATs
  —number of years of math, and
  —high school grades.

- Including more interaction terms would presumably have resulted in a better feel for the joint effects of coaching and other important variables or the conditions under which coaching is most effective.

- Separate analyses for each coaching school, and/or more sophisticated procedures for pooling data, could have been conducted, given the acknowledged differences between the students attending the two coaching schools. Important effects may have been obscured by merely combining together data for all students at both schools as in the demographic profiles.

Hypothesis Testing

At times the FTC study is too selective in the testing of plausible hypotheses. This defect is especially evident in the analyses of self-selection bias.

- As one example, on page 18 of the BCP report, the hypothesis is presented that students who are coached before they take the
SAT for the first time are serious and do not plan to take it a second time, while those who are not coached are not as serious because they do plan to take it a second time. This hypothesis is tested by showing that coached students were more likely than their uncoached counterparts to take the SAT a second time and that, therefore, the issue was not a cause for concern. However, no consideration is given to the counter-hypothesis that because coached students were more likely than uncoached students to take the exam a second time, they are more serious and more determined to persist until they obtain satisfactory scores.

- The FTC study acknowledges the potential for biased estimates of effects resulting from student self-selection. However, the study focuses on only one possible type of self-selection bias, i.e., that students choose to attend coaching schools on the basis of how their first score (SAT or PSAT) compares with what might be expected from their high school rank, their parents' income, etc. This hypothesis is tested by regressing first SAT score (and PSAT score) on demographic variables and showing that coached students, on the average, obtained slightly lower than expected SAT and PSAT scores, whereas noncoached students did slightly better than expected. However, similar analyses using each of the demographic characteristics (instead of PSAT and SAT scores) as dependent variables might very well have shown similar, or even more dramatic, differences between coached and uncoached students, suggesting self-selection on other variables.

One plausible alternative hypothesis is that self-selection operates primarily on the basis of family income, an hypothesis that is supported by the demographic profiles which show a very strong relationship between attending coaching schools and family income. It seems highly likely that regressing family income on other demographic characteristics and PSAT scores would show that coached students come from families whose incomes are higher than expected from other demographic characteristics, and that, therefore, self-selection might well be based primarily on parental income (i.e., on one's family's ability and willingness to pay for coaching). The same analyses could be applied to any of the other available student characteristics.

The following observations suggest that self-selection on the basis of family income is not only consistent with the FTC data
but also at least equally plausible to self-selection based on expectations for PSAT (SAT) scores:

(1) One wonders how precisely students can estimate, or develop strong expectations for, how well they should perform on the PSAT or SAT according to their personal demographic profiles, i.e., how closely statistical expectations correspond to student expectations. As an example, would a student having a particular demographic profile (white male ranking in the top tenth of his class, from a family earning $21,000 a year, having taken four years of math and obtaining an excellent grade in his latest math course) expect to get a PSAT-M score of 56, as would be predicted from the FTC-generated regression equations? How likely would he be to seek coaching if his actual score was below this, or to forego coaching if his actual score exceeded 56?

(2) Differences between coached and noncoached students on variables other than family income could be explained on the basis of incidental selection. If students were to base their decisions to attend coaching schools solely (explicitly) on their parents' ability to pay, differences on most other demographic variables (e.g., PSAT scores, grades, attendance at private schools) could be merely incidental because of the positive correlation of each of these variables with parental income.

(3) Differences in self-selection were detected for students attending the two coaching schools in the FTC study. Self-selection of the type hypothesized in the study was noted for School A, which charges $225, but not for School B, which charges only $75. This pattern is consistent with the alternative hypothesis that family income is the major self-selection variable.

Testing other types of self-selection bias seems very important in light of the FTC conclusion that coaching is effective for "underachievers," which is defined in the report as being the same as self-selection.

Logic and Inferences

Several of the conclusions drawn in the FTC report are open to further consideration.
• The analysis of self-selection bias shows that coached students had unexpectedly low scores on both a previous SAT and the PSAT, from which it is concluded that students who tend to enroll in coaching schools are those who do worse than expected on standardized tests. The fact that, on the average, those choosing coaching schools do worse than expected on both previous exams (SAT and PSAT) is used to infer that self-selection is not on the basis of scoring lower than expected by chance (i.e., that because the phenomenon is noted for two tests, it is not by chance).

However, this analysis rules out only that self-selection is not based solely on obtaining by chance an initial SAT score that was lower than expected. It does not rule out the hypothesis that self-selection is based on scoring low by chance twice (on both the PSAT and the first SAT). That is, those students who are unlucky twice may be the ones choosing to select coaching. Using the FRC technique, we can reason that if, say, about one-fifth of all students tend to be unlucky, i.e., score below expectations on each exam, then about one in twenty-five would be expected to be unlucky both times. Given the relatively small percentage of all examinees who attend commercial coaching schools, coached students could very well come primarily from this doubly unlucky group.

• A major conclusion of the study that rests on the analysis of self-selection is that "coaching can be effective for those who do not score well on standardized tests" (p. 35), and that "if only underachievers can be helped, it is possible that coaching at the second school would be effective for such students" (executive summary). Because of its policy implications, this conclusion deserves critical analysis. The reasoning from which this conclusion results is faulty because self-selection of the test underachievement type was noted for only one of the two schools. That is, there are underachievers at one school but not at the other. This confounding of particular school with type of student does not allow the type of conclusion made in the FRC report. It is not possible to conclude from the FRC analysis of only two schools whether the nature of the program or the type of student in attendance is responsible for the apparent effect on test scores. An analysis of the effects for under- and overachievers within schools is needed to make an inference of this nature.

• Related to the major conclusions listed above is the extremely
weak argument that because self-selection was not detected at School A for one exam [the 1975 Verbal exam, for which coaching was effective], then coaching at School A can be effective for all students, not just underachievers. Not only is this inference not justified, but the reasonableness of the finding upon which it is based [i.e., that self-selection was not detected for one portion (Verbal) of one exam (the 1975 administration) can be questioned]. It is difficult to understand how or why self-selection should be different in one year for only one part of the exam.

Findings

A number of the report's findings, while possibly correct, can be classified as either counterintuitive, curious, or needing further explanation or clarification.

• The finding that the effects of coaching on the Verbal part of the SAT were as large as [in some instances larger than] the effects on the Math section is counterintuitive to what we know about the development of verbal and math abilities as measured by the SAT. This somewhat unexpected finding could have been given further attention in the report.

• Table 3 on page 20 of the FTC report shows two estimates of effects that are significantly different statistically for subsamples, but these differences are not mentioned in the report. The first is the difference between estimates for Verbal effects for subsamples 1 and 2 for School A, suggesting that School A was more effective in 1976 than in 1975. Since School A was well established before 1975, this difference may deserve some attention.

  The second significant difference is between subsamples 5 and 6 for Math at School A, suggesting that School A's instruction is more effective for second-time SAT takers than for first-time SAT takers. Possibly, these significant differences merely reflect the relatively large number of significance tests. However, at least the very large difference in year-to-year effects on Verbal scores for School A should have received some mention. Or should we believe that such fluctuations are typical of this kind of study?

• As mentioned above, the analysis of self-selection discussed in the final section of the report concludes that self-selection was
Critical Notes on the FTC Coaching Study

DONALD A. ROCK, December 17, 1979

The FTC analysis [Boston Memorandum] and reanalysis [Washington Report] of the effect of commercial coaching on admissions test score gains applied standard ANCOVA procedures in an attempt to statistically equate two different populations. The FTC report itself documented that the two populations differ on both demographic and ability variables. Conclusions based on such analyses are only valid if: [1] strong assumptions can be made about the appropriateness of the statistical correction used in eliminating the measured pre-existing differences, and [2] all unmeasured pre-existing differences are unrelated to observed differential gains over the time periods in question. Subsequent reanalyses at the Educational Testing Service suggest that [1] and [2] above are likely to be untrue in some cases and only partially true in others. Some of the failures to meet assumptions were empirically demonstrated (Appendix 3) and their consequences with respect to interpretations outlined. Problems with the FTC analysis and reanalysis, some statistical and some logical, are as follows:

[1] The FTC authors report that approximately 600 scores for coaching school enrollees were not found in the SAT files. This, I think, has serious consequences for their conclusion. If these 600 individuals cancelled their scores because they did not think they did well, then one would expect the estimates of coaching effects to be biased upwards.

[2] The ETS reanalysis of test scores at three points in time suggests that the standard ANCOVA approach used by the FTC was
inappropriate for the verbal data because of self-selection effects which were at least partially captured in differential group growth rates. More appropriate growth related adjustment models yielded verbal coaching effects about one-half the size reported by the FTC report.

The mathematics data were found to be more consistent with the standard ANCOVA model and thus more likely to yield reasonable estimates of coaching effects given the available control variables. This is not to say that the FTC and the ETS re-analysis estimates of the mathematics coaching effect are not overestimates, since the only self-selection causes that have been controlled for were those reflected in differential growth rates and/or available demographics. One missing piece of information is the reliability of the pretest scores. When individuals self-select to treatments, commonly held psychometric wisdom suggests that the standard ANCOVA model can be expected to underadjust to the extent that the pretests are less than perfectly reliable, thereby yielding an overestimate of the coaching effect.

(3) The FTC concludes that the students...

who attended the most ‘‘effective’’ school tended to be underachievers on standardized exams, i.e., they scored lower on standardized exams than would have been predicted given their personal and demographic characteristics (including such factors as grades in school and class rank). If this underachieving was random rather than systematic, the results showing the benefits of the coaching received at School A might have been overstated. Analysis was conducted, however, showing that the underachievement by the students was not due to chance, and probably would have continued in the absence of coaching.

[Executive Summary]

The above conclusions were based on examining the size and direction of the deviations from the weighted pooled regression line. The regression slope and intercepts for this pooled regression line were primarily determined by the noncoached population since the ratio of noncoached to coached individuals is approximately 4:1 in this sample. The FTC report of larger mean deviations for the coached group is therefore not unexpected. The directions or sign of the deviations would depend on the way in which the regression hyperplanes differ. The FTC analysts do not appear to have tested the critical assumptions of homogeneity of the within population regressions before they carried out and interpreted the difference in mean residuals. It would appear that
any conclusions about coaching as working best for "under-achievers" is not warranted based on this type of analysis.

A more satisfactory, yet still inconclusive, analysis would be centered on the coached population only, and would include the regression of the first SAT on the PSAT score and significant demographics as control variables. This analysis would be repeated for the second SAT. That is, the second SAT would be regressed on the first SAT and control demographics. The comparison of the relative sizes of the regression weights associated with the PSAT and the first SAT would indicate if a particular subgroup of the coached population benefits more from coaching. The problem here is that this comparison would only be valid if one first shows that the growth rates of subsets of the coached population are the same in the absence of intervention.

* e.g., If the raw score regression weight is greater than 1.0 for the PSAT but less than 1.0 for the first SAT, then one could infer that when background variables are controlled, low scoring individuals gain more from coaching. Since we don't have reliability estimates for the PSAT and SAT for this population, such conclusions would have to be tentative.
Appendix 2
Stroud Report

Reanalysis of the Federal Trade Commission Study of Commercial Coaching for the SAT

T.W.F. Stroud

Abstract

SAT scores of students attending three commercial coaching schools are compared with those of uncoached students by means of multiple regression techniques. The techniques differ from those used in the Federal Trade Commission (FTC) reanalysis in a number of ways which are described in this report. One essential difference is that the SAT scores were predicted using a multiple regression equation based on uncoached students rather than pooled across coached and uncoached students. Residuals from predicted values based on this equation were then computed for coached students; these residuals were averaged within peak test administration months within coaching schools to produce coaching/self-selection effects. Bayesian methods, utilizing the concept of "borrowing strength" or "smoothing" to obtain estimates for smaller schools, were employed to estimate the overall effect at each of the three schools, averaged out over years.

The general impression given by this analysis is similar in spirit to that given by the FTC reanalysis: for the school with the strongest coaching/self-selection effects, smoothed estimates of the effects averaged over years range from 16.9 points on SAT-Math for Juniors to 28.5 points on SAT-Verbal for juniors, with standard errors in the range 9.2 to 18.6; for other schools the smoothed estimates of the effects tend to be smaller.

In exploring interactions between the coaching/self-selection effect and various background variables, it was discovered that at one school showing overall negligible effects black students exhibited a much higher average coaching/self-selection effect (46 points) than whites. This interaction appeared to be separate from an interaction with parental income, which was also found to be statistically significant.
Summary of Background

This report is a reanalysis of data which were first analyzed by the
Boston Regional Office of the Federal Trade Commission (FTC) (1978) and then reanalyzed by the Bureau of Consumer Protection (BCP) (1979) in the FTC central office in Washington, D.C. The purpose of the original study was to determine how much effect preparation under the guidance of commercial coaching schools had on the Scholastic Aptitude Test (SAT) scores and the Law School Admissions Test (LSAT) scores of the clients of those schools. The reanalysis of the SAT data was undertaken by the BCP because serious defects had been discovered in the original Boston office analysis; specifically, observational studies had been made on two groups, coached and uncoached students, and regression lines of SAT score on Preliminary Scholastic Aptitude Test (PSAT) score had been compared with no attempt either to present standard errors or to correct for values of background variables. There are, in fact, systematic differences between coached and uncoached students with respect to such background characteristics as high school achievement, race, and self-reported parental income. There are also systematic differences in the distribution of these variables for students attending the two schools included in the FTC (1979) reanalysis.

The purpose of the present study is an independent investigation of the SAT data analyzed and reanalyzed by the FTC. Extensive use is made here of background variables, as was done in the FTC reanalysis. However, this reanalysis differs from the FTC reanalysis in a number of ways. These ways will be described, following a summary of the FTC reanalysis.

The FTC data set began with about 2000 enrollees at two coaching schools in the metropolitan New York area during the period 1974-77 for whom SAT scores and background data were available. There were 600 enrollees whose names were not found in the ETS file of test-takers in certain zip code areas of Connecticut, New Jersey, and New York which were considered to be primary market areas for the participating coaching schools. Most of these, presumably, had not previously and did not subsequently take the SAT or did not take it in the metropolitan New York area. There was also a third coaching school whose enrollees' data were not analyzed by the FTC, because the number of individuals involved was considered too small. A control group of about 2500 uncoached students was chosen from among persons taking the
SAT during the same three-year period in the same geographical area using a systematic sample of every 150th student in the file provided by Educational Testing Service (ETS).

In the FTC reanalysis, multiple regression analyses were performed on six subsamples:

1. High school juniors taking the SAT for the first time in April 1975. [76 coached and 607 uncoached students]

2. High school juniors taking the SAT for the first time in April 1976. [247 coached and 617 uncoached students]

3. High school seniors taking the SAT for the second time in November 1975. [98 coached and 396 uncoached students]

4. High school seniors taking the SAT for the second time in November 1976. [177 coached and 387 uncoached students]

5. All high school students taking the SAT for the first time on any of the test dates over the 3-year period. [417 coached and 1763 uncoached students]

6. All high school students taking the SAT for the second time on any of the test dates over the 3-year period. [316 coached and 1267 uncoached students]

Background variables included in all regression analyses, in roughly decreasing order of importance, were: pretest score (PSAT Verbal or Math when SAT1, the first SAT taken, was being predicted, SAT1-Verbal or -Math when SAT2, the second SAT taken, was being predicted, using the Verbal pretest to predict the Verbal score and the Math pretest to predict the Math score), self-reported grade in English or math, self-reported rank-in-class, self-reported years of high school instruction in English or math, self-reported parental income, sex, race, high school type, and number of PSATS taken (for predicting SAT1). Coaching school (School A or B) was entered as a dummy variable. Time between pretest and test was also entered, but it did not increase prediction significantly. Only those students with complete data were entered in the regression analyses.

The analyses showed that students from School A scored significantly higher, on the average, than uncoached students. Ninety-five percent confidence limits for the differences in adjusted means, based on the median lower confidence limit and median upper confidence limit over the 12 analyses (6 subsamples, Verbal
and Math scores), were 14 and 38. Coached students from School B did not earn significantly higher scores than uncoached students, with median confidence limits of -12 and +19.

The FTC (1979) report then presented a regression analysis predicting PSAT scores from the same background variables as above. When actual and predicted PSAT scores were compared, the students coached between the PSAT and SAT1 had significantly lower residual scores on the PSAT than the uncoached students. They thus appeared to be "underachievers on standardized tests." A similar analysis was carried out for SAT1 scores. The finding was that students coached between SAT1 and SAT2 had significantly lower residual scores on SAT1 than uncoached students. The conclusion, as reported in the FTC reanalysis, was that obtaining pre-test scores that were lower than predicted from background variables might have been a factor in enrollees selecting themselves for coaching schools. Since pretest was the strongest predictor variable in the coaching regression analysis, this self-selection effect could account for some of the difference between adjusted means of the coached and uncoached groups.

The FTC reanalysis then attempted to control for this self-selection by recomputing the original regression analysis with the pretest removed from the set of covariates. In this analysis the adjusted means for coached students at School A were above those of uncoached students, but the difference was not significant at the 5% level (median confidence limits of -5 and +30). Finally, in an effort to see whether this self-selection in the form of underachievement on standardized tests was consistent or random, both PSAT and SAT1 scores were predicted for students who were coached between SAT1 and SAT2. Since residual scores for both PSAT and SAT1 for these coached students were found to be negative and significantly lower than residual scores for uncoached students, this was taken as evidence that the underachievement was "consistent" rather than "random." Hence, when the scores of these students improved after attendance at coaching school, the improvement was attributed to coaching rather than to any self-selection effect.

It seems quite plausible, however, especially since the coached groups are so small, that students may have selected themselves for coaching on the basis of negative measurement errors in both PSAT and SAT1. In any case, we are dealing with an observational study where the self-selection effects are likely quite complex and it is not possible, from the data provided, to disentangle the effects of self-selection from the effects of coaching. How much is
a coaching effect and how much is a self-selection effect may be speculated upon, but it cannot be determined from the data.

Differences Between This Analysis and the FTC Reanalysis

In the FTC (1979) reanalysis, verbal background variables (e.g., PSAT-V, grades in English, years of English) were used to predict SAT-Verbal scores and quantitative background variables were used to predict SAT-Math scores. It was found, however, that both Verbal and Math PSAT scores were important in predicting both SAT1-V and SAT1-M. Similarly, both SAT1-V and -M scores were valuable in predicting both SAT2-V and SAT2-M for the seniors. The number of years of English taken did not additionally contribute to the predictive value of any regression equation when the other variables were already entered.

The FTC reanalysis dealt with the six subsamples given earlier. The analysis presented here deals with the peak 1975 month and peak 1976 month only—subsamples 1, 2, 3, and 4 in the FTC study—and omits the combined subsamples 5 and 6 which were regarded as too heterogeneous.

An additional difference is that missing-value techniques were used here so that students who did not report parental income, race, or rank-in-class could nevertheless be included in the present analysis; such students were excluded from the regression equations in the FTC reanalysis. Furthermore, in the FTC reanalysis, students from the coaching schools who did not receive their coaching prior to the SAT administration in question were added to the sample of uncoached students. Since this destroys the representativeness of the systematic sample of every 150th uncoached student, they were excluded in the present analysis.

Although the FTC's data set involved three coaching schools, only Schools A and B were used in the FTC reanalysis, because the enrollment in School C was considered to be too small. Furthermore, the student populations from Schools A and B were frequently amalgamated so that the report refers to "coached students." The position taken here is that each coaching school is a separate population and thus the three coaching schools are identified separately. Estimates are provided for these three schools which utilize the empirical Bayes concept of borrowing strength, and which also allow for the possibility of predicting coaching/self-selection effects in the same schools in future years (under certain assumptions).
Aims of This Analysis and Summary of Statistical Methods

Students who select a coaching program at a commercial coaching school tend to differ in two major ways from the vast majority who do not do so, namely, motivation and financial means. Those who do not have higher-than-average financial means must be even more strongly motivated. Although there are many students who do not obtain coaching who are highly motivated to do well on the SAT, there are also many whose motivation is poor. A reduced motivation was concluded to be one of the factors involved in the SAT score decline of recent years (Wirtz et al., 1977).

The view taken in the present analysis is that the data provide no way of distinguishing the effect of coaching from the effect of self-selection or motivation. There is a rough proxy for financial means (self-reported parental income) but there is no proxy for motivation. There may also be other factors in the self-selection into coaching which we have not pinpointed. Therefore, differences in SAT score after coaching between coached students and uncoached students will be called combined coaching/self-selection effects.

Because of the obvious differences in character between the coached and uncoached populations (e.g., higher grades and rank-in-class, higher parental income, and higher percent enrollment in private schools for coached students), it would be unreasonable to suppose that the mean SAT scores, after controlling for all covariates imaginable, would be exactly the same. Hence, an hypothesis of no coaching/self-selection effect was avoided. Instead, the coaching/self-selection effect at each coaching school was estimated along with a standard error for the estimate. A coaching/self-selection effect whose magnitude exceeds a certain number of standard errors (say 2 or 2.5) would be regarded as being "significant" in the traditional sense.

The analyses contained herein were performed separately on four dependent variables: SAT1-V scores of juniors, SAT1-M scores of juniors, SAT2-V scores of seniors, and SAT2-M scores of seniors. As indicated earlier, analysis was restricted to students whose SAT was taken in the peak month (April for juniors, November for seniors) of 1975 and 1976.

The first stage of the analysis was to obtain a multiple regression equation for juniors and a multiple regression equation for seniors for each dependent variable based on the sample of uncoached students. Next, these equations were applied to the coached students in order to predict the SAT scores for the coached
students as if they had been uncoached students with the same values of predictor variables. Residuals from these regressions were calculated. The coaching/self-selection effect is thought of as the typical value of these residuals. Separate coaching/self-selection effects, together with standard errors, were calculated for each of the two administration years for each of the three coaching schools. The third and final stage of the analysis was an empirical Bayes smoothing of the effects across administration years and coaching schools. The analysis is described in greater detail in the following three sections.

The basic aim of these analyses was to estimate the magnitude of the coaching/self-selection effects. If these combined effects are small, this puts a ceiling on the effects due to coaching at the three schools considered. If these combined effects are large, one can only speculate on how much of the effects can be attributed to coaching and how much to self-selection.

\[ Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \quad , \quad i = 1, 2, \ldots, N \]  \quad \{1\}

where \( Y_i \) represents the score on the dependent variable \( \text{SAT1-V}, \text{SAT1-M}, \text{SAT2-V}, \text{or SAT2-M} \) of the \( i \)th student (junior or senior, depending on the dependent variable), \( N \) is the number of uncoached students, and \( X_{ij} \) represents the score on the \( j \)th predictor variable of the \( i \)th student. The set of \( p \) predictor variables includes a dummy variable representing the administration year. The error \( \epsilon_i \) is assumed to have constant variance \( \sigma^2 \).

Although the \( Y_i \) were measured for all students, some students had \( X_{ij} \) missing for one or more \( j \). Thus, instead of obtaining the coefficients \( \beta_j \) in equation \( \{1\} \) from a least-squares regression program, the following method was used. The maximum-likelihood estimates of the means, variances, and covariances of \( \{Y_i, X_{i1}, X_{i2}, \ldots, X_{ip}\} \) were computed using the program based on a multivariate normal model for all variables with some observations missing (Rubin, 1974). Estimates of the coefficients \( \beta_0 \) and \( \beta_j \) (\( j = 1, \ldots, p \)) were obtained from the resulting estimates using the sweep operator \( \{\text{swp}\} \) on all predictor variables. If all variables were observed on all individuals, the estimated \( \beta \)s so obtained would be the least-squares estimates. This method is more effi-
cient than that of computing covariances using pairwise-observed data and provides consistent estimates under more modest conditions (Rubin, 1976).

Regression coefficients were also computed for use with coached students who had one or more of the predictor variables of race, high school rank, and parental income missing. For example, for a student who had all three of these variables missing a regression equation which did not involve these three predictors was used. To obtain this equation, one begins with the matrix of maximum-likelihood estimates referred to in the preceding paragraph and applies the svp operator to the set of all predictor variables excluding race, rank, and parental income. This yields the regression coefficients presented in the last column of Tables 1a to 1d, which show all regression coefficients for predicting sat1-V, sat1-M, sat2-V, and sat2-M, respectively. Columns 2 to 7 give regression coefficients for other missing variable patterns, and column 1 gives the complete-data regression. The list of all predictor variables studied appears in Table 1e.

**Estimates and Standard Errors of the Coaching/Self-Selection Effect**

The next step is to predict the sat scores of coached students using the equation (1) which was developed for uncoached students. Denote by $b_0, b_1, \ldots, b_p$ the estimates of $\beta_0, \beta_1, \ldots, \beta_p$ obtained from that analysis. Define

$$\hat{Y}_{sti} = b_0 + \sum_{j=1}^{p} b_j X_{stij}$$

(2)

for the $ith$ student from coaching school $s$ taking the sat in year $t$ whose background data $\{X_{sti}, X_{stij}, \ldots, X_{stip}\}$ is complete. For a student with missing data, background variables comprise a smaller set of predictor variables having subscripts in the subset $A \subset \{1, 2, \ldots, p\}$ (and whose scores are missing in the complementary set $A^c$), define

$$\hat{Y}_{sti} = b_0^A + \sum_{j \in A} b_j^A X_{stij}$$

(3)

where $b_j^A$ are the regression coefficient estimates based on an equation using only the variables in the set $A$. The values of $b_j$ and $b_j^A$ for the subsets $A$ being considered are the values presented in Table 1a to 1e. $Y_{sti}$ represents the predicted score of the $(s,t,i)th$ coached student, ignoring the fact that he was coached.

For each coached student, the quantity $d_{sti} = Y_{sti} - \hat{Y}_{sti}$ is then
computed. This is the residual of his actual score from his predicted score and represents the estimated coaching/self-selection effect for that one student. The average of these residuals within the \{s,t\} cell, denoted by \(d_{st}\), is an unbiased estimate of the parameter \(\tilde{d}_{st}\), which is defined as the average difference between the value of \(E(Y|X)\) under a "correct" model for coached students in the \{s,t\} cell (which may be a different value for each student in the cell) and the value of \(E(Y|X)\) under the uncoached model \(\{1\}\). If one assumed a strictly additive coaching/self-selection effect according to the model

\[
Y_{sti} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{sti} + \delta_{st} + \epsilon_{sti},
\]

where \(\beta_0, \beta_1, \ldots, \beta_p\) are the same parameters as in the uncoached model, then it would be better to use, instead of \(d_{st}\), a weighted average which would be minimum-variance unbiased under this model. However, as pointed out in Cochran and Rubin (1973, p. 424), \(d_{st}\) is an unbiased estimate of \(\tilde{d}_{st}\) even when \(\beta_0, \beta_1, \ldots, \beta_p\) take on different values in the coached model.

Tables 2 and 3 show the sample sizes for the uncoached students and the coached students, respectively. Table 3 also shows the number of students who were removed from the analysis of coached students because of rare or extensive missing data patterns. To have included these data patterns in the analysis would have made the computing task much more onerous and was not deemed to be worthwhile.

Table 4 shows the estimated coaching/self-selection effects \(\tilde{d}_{st}\), together with estimated standard errors. These standard errors are calculated on the assumption that model \(\{3\}\) holds for the coached students.

**Smoothed Estimates for the Three Coaching Schools and Predictive Distributions for a Future Year**

Bayesian methods are recommended to predict the coaching/self-selection effect in a future year. The strongest advantage of Bayesian prediction is that a predictive distribution for a future value can be obtained, which can be made to incorporate uncertainty due to the effects of sampling in the observed data, uncertainty due to year-to-year fluctuation, and uncertainty due to not knowing certain quantities such as variances between schools and variances between years. An additional advantage is that information from schools with large numbers of students can be
used to improve the estimates for schools with smaller numbers of students. (In problems involving larger numbers of groups, a general improvement can be brought about by smoothing all predictions toward the grand mean.) In this section the Bayesian analysis is incorporated into the model and the smoothed estimates of the Verbal and Math coaching/self-selection effects for each school averaged over years is presented. These smoothed estimates also serve as point predictors for a future year. Standard errors for these predictions are also shown. The validity of these predictions and standard errors depend upon the following assumptions:

1. The observed $\tilde{d}_{st}$ are unbiased estimates of the coaching/self-selection effects $\delta_{st}$; the standard errors shown in Table 4 represent the uncertainty of these estimates.

2. The quantities $\tilde{\delta}_s$ within school s for different t, both past and future, are normally distributed about the expected effect $\delta^*_s$ for that school, which remains constant within the future range being predicted. The variance of the $\tilde{\delta}_s$ about $\delta^*_s$ is the same for each school s. The value being predicted is $\tilde{\delta}_s$ where t represents a future year.

In the Bayesian analysis, the prior distribution of the expected effects $\delta^*_s$ is a normal distribution. The standard deviation of this distribution and the standard deviation of the $\tilde{\delta}_s$ within school have independent prior distributions which are each approximately constant over the range of plausible values. The posterior means and standard deviations of the $\delta^*_s$ are shown in Table 5. The means and standard deviations of the predictive distributions for each $\tilde{\delta}_s$ (s = A, B, C; t = a future year) for the four dependent variables (juniors' sat1-V and sat1-M, seniors' sat2-V and sat2-M) are shown in Table 6.

*Interactions between Coaching/Self-Selection Effects and Background Variables*

The question now arises: is the coaching/self-selection effect greater for some subsets of the student population than for others? To explore possible evidence of such interactions, a multiple regression analysis of the Verbal and of the Math individual coaching/self-selection effects was performed on the set of variables which had been used to predict the sat scores. This was done for the three largest cells of coached students, namely the
SAT2 seniors of School A and the SAT1 juniors of Schools A and B, all from the 1976 peak month administration. Only students with no missing values on background variables were entered into this exploratory analysis; the sample sizes were 102, 103, and 85, respectively. For the two School A cells, the sums of squares and degrees of freedom for multiple regression and for residual are shown in Table 7; the F-values of 1.106 (Verbal) and 1.119 (Math) for seniors and the F-values of 0.606 (Verbal) and 0.735 (Math) for juniors are not significant.

For the School B juniors, the sums of squares and degrees of freedom for multiple regression and for residual are also shown in Table 7; the F-value of 2.112 (Verbal) is significant at the .05 level, while the F-value of 1.587 (Math) is not. To discover which variables contributed most to the analysis of the Verbal individual coaching/self-selection effects, the squared correlation coefficient \( r^2 \) was calculated between this dependent variable and each of the predictor variables; these are shown in Table 8. The most significant predictor variable is RACE2 (= 1 for black, 0 for nonblack) with \( r^2 = .0830 \). The square of the partial correlation coefficient \( r^2 \text{RACE}_2 \) was calculated between every other predictor variable and the Verbal individual coaching/self-selection effect and is also shown in Table 8. The most important predictor variable after RACE2 is entered is INCOME, with \( r^2 \text{RACE}_2 = .0664 \), or \( r \text{RACE}_2 = -.258 \). This differs little from the \( r^2 = .0626 \) \( r = -.250 \) ignoring RACE2, indicating that the effect of self-reported parental income and the effect of black vs. nonblack are essentially separate from each other. Both these effects are statistically significant at the .05 level.

The remainder of this section is devoted to the study of differences in the coaching/self-selection effect across races or minority groups.

Table 9 shows values of background variables and of the individual coaching/self-selection effects for coached students who identified themselves as blacks who took SAT1 as a junior or who took SAT2 as a senior during the peak administration month (and who were thus included in the analysis of this report). There is a strong tendency for both the Verbal and Math coaching/self-selection effects to be above the average (cf. Table 4). For the Verbal effects the pattern is striking: thirteen out of fifteen effects are above average and the remaining two are only very slightly below average.

A significance test was performed between the blacks and whites from Coaching School B, since 13 out of 15 coached blacks
were coached at that school. The variable studied was the Verbal individual coaching/self-selection effect; the null hypothesis was that the mean for blacks equals the mean for whites. Since the overall Verbal averages for the three School B cells (1976 juniors, 1975 seniors, 1976 seniors) were within a fairly narrow range (4.16 to 11.12, from Table 4), the students from these cells were pooled.

The results of this black-white comparison are shown in Table 10. (The data set used excludes 18 students who identified themselves as belonging to other nonblack minority groups and who are among those listed in Table 12.) Equality of population variances for black and white coaching/self-selection effects was not assumed. The Welch approximate t-test was used (see, e.g., Brownlee, 1965, p. 299). The 2-sided test showed significance at the .001 level, with the average for blacks being 46.7 points above that for whites. The average Math coaching/self-selection effect for blacks also exceeded that for whites, but the difference was not statistically significant.

There were five black juniors taking SAT1 and one black senior taking SAT2 at times other than peak administration months, so their scores have not been included in any analysis described so far. A tabulation of their data appears in Table 11. The Verbal effects are higher than the corresponding cell averages in five cases out of six; the average of these six Verbal effects is 52.6 points above the average of the corresponding cell average values. The Math effects are higher than the cell averages four times out of six, averaging 13.6 points above cell average. This pattern is virtually the same as that observed for the blacks who took the SAT in the peak administration months.

Finally, the data for students who identified themselves as belonging to minority groups other than black are presented in Table 12. Both peak-month and nonpeak-month SAT-takers are included. The individual coaching/self-selection effects bear more resemblance to those of white students than to those of black students.

Conclusions

The year-by-year estimated effects for coaching schools A and B, as given in Table 4, are of a similar size to those presented in the FTC report (1979), in that for School A the coaching/self-selection effects range from 20 to 34 SAT points (from 3 to 6.9 standard
errors) and for School B the effects are nonsignificant except for the seniors' sat2-M in 1975 [26 points = 2.38 standard errors]. For School C the estimates fluctuate markedly from year to year, the strongest effect being the juniors' sat1-M in 1975, +2.2 standard errors as compared with −1.5 standard errors in 1976.

The smoothed estimates of the school-by-school effects averaged over years, given by Table 5, are only slightly smaller for School A than the unsmoothed estimates. However, the standard deviations shown in Table 5 are larger than the within-year standard errors of Table 4, reflecting the year-to-year fluctuation exhibited in the data. For a future year one would predict a positive coaching/self-selection effect at all schools (Table 6), but with a standard error greater than the value of the predicted effect except for sat-V at School A (predicted effect = 1.8 standard errors for juniors, 1.3 for seniors) and sat-V for juniors at School C (predicted effect = 1.3 standard errors).

It is impossible to judge the extent to which each of these effects is due to the coaching program or rather to personal characteristics underlying self-selection of the coached students. However, in a randomized experiment (Alderman and Powers, 1979; Rubin, 1979) which was designed to be free of self-selection effects, special preparation programs offered in eight high schools had estimated coaching effects ranging from 6.3 to 12.5 sat points, or 0.9 to 1.9 standard errors. The larger estimates derived from the data analyzed in this report are consistent with the possibility that both the motivation to do well on tests, as exemplified by enrollment in a commercial coaching school, and the coaching program itself contribute to a small gain in sat scores.

There is a clear and strong positive difference in the coaching/self-selection effect of blacks over whites at School B, as shown by Table 10. Whereas the average effect for the 132 whites analyzed here was virtually nil, the average coaching/self-selection effect for the 13 blacks was 46 points, or 4.47 standard errors \(|S_x/\sqrt{N} = 10.39\). One possible interpretation of this phenomenon is that blacks have an initial disadvantage in test-taking that can be overcome by coaching. However, it is also possible that there is a greater difference in personal factors such as motivation between blacks who enroll in a coaching school and blacks who do not than between whites who enroll and whites who do not. We caution that it would be unwarranted to take the relative unimportance of the coefficients for race in the multiple regressions (Table 1) as evidence against the hypothesis that blacks are disadvantaged in test-taking. The strongest predictor variable in the
multiple regression is the pretest [PSAT or SAT], and a race effect could conceivably be operating there.

To summarize, this study has shown certain relatively small but unmistakable effects of coaching and/or of self-selection. Further study is needed, perhaps in the form of additional randomized experiments, to shed more light on the subject of what portion of these effects is due to coaching.
### Table 1a
Regression Coefficients for Predicting sat1-V Scores for Juniors
[April Administration]
For Complete Data and Various Combinations of Missing Variables
[For names of predictor variables see Table 1e]

<table>
<thead>
<tr>
<th>Variable No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.37</td>
<td>-.19</td>
<td>.33</td>
<td>-.23</td>
<td>.99</td>
<td>.44</td>
<td>.97</td>
<td>.41</td>
</tr>
<tr>
<td>2</td>
<td>1.692</td>
<td>1.908</td>
<td>1.680</td>
<td>1.896</td>
<td>1.745</td>
<td>1.958</td>
<td>1.730</td>
<td>1.945</td>
</tr>
<tr>
<td>3</td>
<td>-.033</td>
<td>.181</td>
<td>-.042</td>
<td>.173</td>
<td>-.052</td>
<td>.161</td>
<td>-.069</td>
<td>.146</td>
</tr>
<tr>
<td>4</td>
<td>-.3.14</td>
<td>-.3.99</td>
<td>-.3.04</td>
<td>-.3.87</td>
<td>-.3.77</td>
<td>-.4.61</td>
<td>-.3.59</td>
<td>-.4.42</td>
</tr>
<tr>
<td>5</td>
<td>.4817</td>
<td>.4786</td>
<td>.4958</td>
<td>.4940</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>-.6.1</td>
<td>-.6.5</td>
<td>—</td>
<td>—</td>
<td>-.8.5</td>
<td>-.8.9</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7</td>
<td>-.1.1</td>
<td>-.1.4</td>
<td>—</td>
<td>—</td>
<td>-.3.8</td>
<td>-.4.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>.3131</td>
<td>—</td>
<td>.3151</td>
<td>—</td>
<td>.3107</td>
<td>—</td>
<td>.3136</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>13.3</td>
<td>14.1</td>
<td>13.4</td>
<td>14.2</td>
<td>13.4</td>
<td>14.2</td>
<td>13.6</td>
<td>14.4</td>
</tr>
<tr>
<td>10</td>
<td>.174</td>
<td>.199</td>
<td>.182</td>
<td>.208</td>
<td>.232</td>
<td>.257</td>
<td>.247</td>
<td>.273</td>
</tr>
<tr>
<td>11</td>
<td>1.093</td>
<td>1.197</td>
<td>1.096</td>
<td>1.201</td>
<td>1.142</td>
<td>1.245</td>
<td>1.153</td>
<td>1.258</td>
</tr>
</tbody>
</table>

### Table 1b
Regression Coefficients for Predicting sat1-M Scores for Juniors
[April Administration]
For Complete Data and Various Combinations of Missing Variables
[For names of predictor variables see Table 1e]

<table>
<thead>
<tr>
<th>Variable No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-61.6</td>
<td>-85.7</td>
<td>-65.0</td>
<td>-88.9</td>
<td>-60.2</td>
<td>-84.2</td>
<td>-63.5</td>
<td>-87.4</td>
</tr>
<tr>
<td>2</td>
<td>.028</td>
<td>.246</td>
<td>.063</td>
<td>.278</td>
<td>.049</td>
<td>.265</td>
<td>.082</td>
<td>.296</td>
</tr>
<tr>
<td>3</td>
<td>1.575</td>
<td>1.792</td>
<td>1.571</td>
<td>1.785</td>
<td>1.567</td>
<td>1.784</td>
<td>1.561</td>
<td>1.775</td>
</tr>
<tr>
<td>4</td>
<td>-.8.67</td>
<td>-.9.52</td>
<td>-.8.39</td>
<td>-.9.23</td>
<td>-.8.92</td>
<td>-.9.76</td>
<td>-.8.60</td>
<td>-.9.43</td>
</tr>
<tr>
<td>5</td>
<td>.1889</td>
<td>.1858</td>
<td>.1870</td>
<td>.1853</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>18.1</td>
<td>17.6</td>
<td>—</td>
<td>—</td>
<td>17.1</td>
<td>16.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7</td>
<td>-16.8</td>
<td>-17.1</td>
<td>—</td>
<td>—</td>
<td>-17.8</td>
<td>-18.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>.3163</td>
<td>—</td>
<td>.3135</td>
<td>—</td>
<td>.3153</td>
<td>—</td>
<td>.3130</td>
<td>—</td>
</tr>
<tr>
<td>12</td>
<td>1.478</td>
<td>1.575</td>
<td>1.474</td>
<td>1.571</td>
<td>1.485</td>
<td>1.582</td>
<td>1.483</td>
<td>1.579</td>
</tr>
</tbody>
</table>

111
### Table 3
Numbers of Coached Juniors With Complete and With Incomplete Data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>49</td>
<td>103</td>
<td>85</td>
<td>16</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Missing:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income, Race, Rank</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Income, Race</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Income, Rank</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>14</td>
<td>19</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Race, Rank</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Total Students Used in Analysis</strong></td>
<td>71</td>
<td>134</td>
<td>108</td>
<td>21</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td><strong>Other Data Patterns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(not used in analysis)</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Total Coached Students with PSAT and some background data</strong></td>
<td>76</td>
<td>138</td>
<td>111</td>
<td>24</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

### Numbers of Coached Seniors With Complete and With Incomplete Data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>45</td>
<td>102</td>
<td>15</td>
<td>24</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Missing:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income, Race, Rank</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Income, Race</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Income, Rank</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>10</td>
<td>13</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Race, Rank</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Total Students Used in Analysis</strong></td>
<td>65</td>
<td>123</td>
<td>21</td>
<td>34</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td><strong>Other Data Patterns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(not used in analysis)</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total Coached Students with PSAT and some background data</strong></td>
<td>70</td>
<td>130</td>
<td>21</td>
<td>34</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 4
Coaching/Self-Selection Effects and Their Standard Errors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Juniors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>27.44</td>
<td>6.66</td>
<td>34.13</td>
<td>5.23</td>
<td>A</td>
<td>22.36</td>
<td>7.45</td>
<td>20.37</td>
</tr>
<tr>
<td>B</td>
<td>—</td>
<td>—</td>
<td>4.16</td>
<td>6.38</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>C</td>
<td>17.88</td>
<td>11.19</td>
<td>25.64</td>
<td>14.10</td>
<td>C</td>
<td>27.68</td>
<td>12.52</td>
<td>—23.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Seniors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>32.07</td>
<td>5.86</td>
<td>31.34</td>
<td>4.51</td>
<td>A</td>
<td>22.54</td>
<td>6.41</td>
<td>32.62</td>
</tr>
<tr>
<td>B</td>
<td>11.12</td>
<td>10.16</td>
<td>4.76</td>
<td>8.05</td>
<td>B</td>
<td>26.40</td>
<td>11.10</td>
<td>8.07</td>
</tr>
<tr>
<td>C</td>
<td>—10.47</td>
<td>14.02</td>
<td>20.30</td>
<td>18.70</td>
<td>C</td>
<td>24.34</td>
<td>15.33</td>
<td>—21.57</td>
</tr>
</tbody>
</table>

Table 5
Smoothed Estimates of Coaching/Self-Selection Effects
(Posterior Means and Standard Deviations)

Juniors Taking SAT1

<table>
<thead>
<tr>
<th>School</th>
<th>Verbal Mean</th>
<th>Std. Deviation</th>
<th>Quantitative Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.5</td>
<td>9.2</td>
<td>16.9</td>
<td>18.6</td>
</tr>
<tr>
<td>B</td>
<td>9.7</td>
<td>12.8</td>
<td>7.8</td>
<td>23.0</td>
</tr>
<tr>
<td>C</td>
<td>20.9</td>
<td>10.7</td>
<td>8.4</td>
<td>19.5</td>
</tr>
</tbody>
</table>

Seniors Taking SAT2

<table>
<thead>
<tr>
<th>School</th>
<th>Verbal Mean</th>
<th>Std. Deviation</th>
<th>Quantitative Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.4</td>
<td>12.1</td>
<td>24.0</td>
<td>15.1</td>
</tr>
<tr>
<td>B</td>
<td>9.4</td>
<td>12.3</td>
<td>17.2</td>
<td>15.3</td>
</tr>
<tr>
<td>C</td>
<td>6.1</td>
<td>15.0</td>
<td>12.0</td>
<td>17.6</td>
</tr>
</tbody>
</table>
### Table 6

Predictive Means and Standard Standard Deviations of Coaching/Self-Selection Effects for a Future Year

<table>
<thead>
<tr>
<th>School</th>
<th>Verbal Mean</th>
<th>Std. Deviation</th>
<th>Quantitative Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.5</td>
<td>15.7</td>
<td>16.9</td>
<td>35.9</td>
</tr>
<tr>
<td>B</td>
<td>9.7</td>
<td>18.0</td>
<td>7.8</td>
<td>38.4</td>
</tr>
<tr>
<td>C</td>
<td>20.9</td>
<td>16.7</td>
<td>8.4</td>
<td>36.4</td>
</tr>
</tbody>
</table>

### Table 7

Multiple Regression Statistics for the Regression of the Individual Coaching/Self-Selection Effect on Pretest and Background Variables

(1976 Peak Month Administration)

<table>
<thead>
<tr>
<th>School A Juniors</th>
<th>School A Seniors</th>
<th>School B Juniors</th>
</tr>
</thead>
<tbody>
<tr>
<td>sat1-V</td>
<td>sat1-Q</td>
<td>sat2-V</td>
</tr>
<tr>
<td>Regression: SS</td>
<td>21396</td>
<td>23353</td>
</tr>
<tr>
<td>Regression: DF</td>
<td>10*</td>
<td>10*</td>
</tr>
<tr>
<td>Residual: SS</td>
<td>324662</td>
<td>292338</td>
</tr>
<tr>
<td>Residual: DF</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.062</td>
<td>.074</td>
</tr>
<tr>
<td>R</td>
<td>.249</td>
<td>.272</td>
</tr>
<tr>
<td>F</td>
<td>.606</td>
<td>.735</td>
</tr>
</tbody>
</table>

*The predictor variable RACE2 (black vs. nonblack) was omitted since there were no black students in this cell.
Table 8
Squared Correlation Coefficients between the Individual Coaching/Self-Selection Effects and Each Predictor Variable: Zero-Order Correlation \( r^2 \) and Partial Correlation after Entering \( RACE2 \) \( r^2_{RACE2} \)

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>( r^2 )</th>
<th>( r^2_{RACE2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported English grade</td>
<td>.0001</td>
<td>.0005</td>
</tr>
<tr>
<td>Self-reported Math grade</td>
<td>.0006</td>
<td>.0002</td>
</tr>
<tr>
<td>High school type</td>
<td>.0598</td>
<td>.0524</td>
</tr>
<tr>
<td>Self-reported parental income</td>
<td>.0626</td>
<td>.0664</td>
</tr>
<tr>
<td>RACE1</td>
<td>.0304</td>
<td>.0429</td>
</tr>
<tr>
<td>( RACE2 )</td>
<td>.0830</td>
<td>—</td>
</tr>
<tr>
<td>Self-reported rank</td>
<td>.0168</td>
<td>.0093</td>
</tr>
<tr>
<td>Sex</td>
<td>.0492</td>
<td>.0433</td>
</tr>
<tr>
<td>Years of mathematics</td>
<td>.0121</td>
<td>.0140</td>
</tr>
<tr>
<td>PSAT-V</td>
<td>.0211</td>
<td>.0282</td>
</tr>
<tr>
<td>PSAT-M</td>
<td>.0001</td>
<td>.0004</td>
</tr>
</tbody>
</table>
Table 9
Background Variables and Individual Coaching/Self-Selection Effects for Coached Students Identified as Black

<table>
<thead>
<tr>
<th>Coaching School</th>
<th>Adm. Date</th>
<th>English Grade</th>
<th>Math Grade</th>
<th>H.S. Type</th>
<th>Potential Income</th>
<th>Rank in Class</th>
<th>Sex</th>
<th>Years of Math</th>
<th>PSAT M</th>
<th>SAT1 V M</th>
<th>SAT2 V M</th>
<th>C/s Effect V M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Juniors</td>
<td>B</td>
<td>4/76</td>
<td>85</td>
<td>NP</td>
<td>7</td>
<td>70</td>
<td>F</td>
<td>4</td>
<td>39</td>
<td>33</td>
<td>390</td>
<td>380</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>85</td>
<td>75</td>
<td>NP</td>
<td>70</td>
<td>F</td>
<td>2</td>
<td>40</td>
<td>42</td>
<td>460</td>
<td>450</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>85</td>
<td>85</td>
<td>NP</td>
<td>5</td>
<td>70</td>
<td>M</td>
<td>4</td>
<td>31</td>
<td>43</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>85</td>
<td>95</td>
<td>P</td>
<td>17</td>
<td>70</td>
<td>M</td>
<td>4</td>
<td>49</td>
<td>49</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>85</td>
<td>85</td>
<td>P</td>
<td>10</td>
<td></td>
<td>M</td>
<td>4</td>
<td>49</td>
<td>36</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>85</td>
<td>95</td>
<td>P</td>
<td>—</td>
<td>70</td>
<td>F</td>
<td>4</td>
<td>33</td>
<td>58</td>
<td>450</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>95</td>
<td>75</td>
<td>NP</td>
<td>35</td>
<td>85</td>
<td>F</td>
<td>2</td>
<td>45</td>
<td>43</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4/76</td>
<td>95</td>
<td>85</td>
<td>NP</td>
<td>35</td>
<td>85</td>
<td>F</td>
<td>3</td>
<td>53</td>
<td>40</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>4/76</td>
<td>85</td>
<td>85</td>
<td>NP</td>
<td>10</td>
<td>70</td>
<td>M</td>
<td>3</td>
<td>31</td>
<td>31</td>
<td>390</td>
</tr>
<tr>
<td>Seniors</td>
<td>A</td>
<td>11/76</td>
<td>85</td>
<td>95</td>
<td>P</td>
<td>29</td>
<td>95</td>
<td>M</td>
<td>4</td>
<td>35</td>
<td>46</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>11/75</td>
<td>85</td>
<td>85</td>
<td>NP</td>
<td>8</td>
<td>—</td>
<td>M</td>
<td>2</td>
<td>38</td>
<td>48</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>11/76</td>
<td>75</td>
<td>85</td>
<td>NP</td>
<td>—</td>
<td>30</td>
<td>F</td>
<td>4</td>
<td>41</td>
<td>38</td>
<td>440</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>11/76</td>
<td>85</td>
<td>85</td>
<td>NP</td>
<td>35</td>
<td>50</td>
<td>M</td>
<td>4</td>
<td>32</td>
<td>34</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>11/76</td>
<td>85</td>
<td>85</td>
<td>NP</td>
<td>5</td>
<td>70</td>
<td>F</td>
<td>4</td>
<td>35</td>
<td>29</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>11/76</td>
<td>85</td>
<td>85</td>
<td>NP</td>
<td>—</td>
<td>50</td>
<td>F</td>
<td>3</td>
<td>33</td>
<td>33</td>
<td>270</td>
</tr>
</tbody>
</table>

*—P = Public, NP = Nonpublic
Table 10
Test of Equality of Mean Verbal Coaching/Self-Selection Effect for Blacks vs. Whites for Coaching School B
(Peak Administration Months, 1975 and 1976, Junior sat1 and Senior sat2 pooled)

<table>
<thead>
<tr>
<th></th>
<th>Blacks x</th>
<th>Whites y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>$N_x = 13$</td>
<td>$N_y = 132$</td>
</tr>
<tr>
<td>Average Effect</td>
<td>$\bar{X} = 46.494$</td>
<td>$\bar{Y} = -.255$</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>$s^2_X = 1404.7$</td>
<td>$s^2_Y = 2370.5$</td>
</tr>
<tr>
<td>$t = \sqrt{\frac{s^2_X}{N_x} + \frac{s^2_Y}{N_y}}$</td>
<td>4.165</td>
<td></td>
</tr>
<tr>
<td>Degrees of Freedom (Welch)</td>
<td>16.28</td>
<td></td>
</tr>
<tr>
<td>P &lt; .001 (two-sided)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_{.0005} = 4.00$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11
Data for Six Black Coached Students Who Took the sat at Times Other Than Peak Administrations

<table>
<thead>
<tr>
<th>Junior (J) or Senior (S)</th>
<th>J</th>
<th>J</th>
<th>J</th>
<th>J</th>
<th>J</th>
<th>J</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administration Date</td>
<td>6/75</td>
<td>12/75</td>
<td>6/76</td>
<td>12/76</td>
<td>6/75</td>
<td>6/75</td>
<td></td>
</tr>
<tr>
<td>Coaching School</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>English Grade</td>
<td>87</td>
<td>85</td>
<td>75</td>
<td>85</td>
<td>95</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>Math Grade</td>
<td>85</td>
<td>95</td>
<td>85</td>
<td>95</td>
<td>85</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>High School Type</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Parental Income</td>
<td>10</td>
<td>19</td>
<td></td>
<td>29</td>
<td>35</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Rank in Class</td>
<td>85</td>
<td>85</td>
<td>70</td>
<td>85</td>
<td>70</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Years of Math</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PSAT-V</td>
<td>29</td>
<td>32</td>
<td>39</td>
<td>33</td>
<td>54</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>PSAT-M</td>
<td>41</td>
<td>41</td>
<td>43</td>
<td>38</td>
<td>47</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>SAT1-V</td>
<td>410</td>
<td>450</td>
<td>400</td>
<td>410</td>
<td>550</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>SAT1-M</td>
<td>490</td>
<td>380</td>
<td>430</td>
<td>470</td>
<td>490</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td>SAT2-V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT2-M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c/ss Effect [V]</td>
<td>82.7</td>
<td>103.4</td>
<td>11.9</td>
<td>51.5</td>
<td>17.2</td>
<td>138.8</td>
<td></td>
</tr>
<tr>
<td>Average c/ss Effect [V]</td>
<td>27.4</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
<td>17.9</td>
<td>32.1</td>
<td></td>
</tr>
<tr>
<td>c/ss Effect [M]</td>
<td>68.9</td>
<td>-45.8</td>
<td>8.1</td>
<td>44.8</td>
<td>-9.4</td>
<td>98.6</td>
<td></td>
</tr>
<tr>
<td>Average c/ss Effect [M]</td>
<td>22.4</td>
<td>3.7</td>
<td>3.7</td>
<td>3.7</td>
<td>27.7</td>
<td>22.5</td>
<td></td>
</tr>
</tbody>
</table>

1 Calculated from regression equation based on nearest Peak Administration Date.
2 Average V- or M-effect for Coaching School at nearest Peak Administration Date.
<table>
<thead>
<tr>
<th>Coaching School Date</th>
<th>English Grade</th>
<th>Math Grade</th>
<th>H.S. Type*</th>
<th>Potential Income</th>
<th>Rank in Class</th>
<th>Sex</th>
<th>Years of Math</th>
<th>V PSAT</th>
<th>V SAT1</th>
<th>V SAT2</th>
<th>V SAT3</th>
<th>C/M Effect</th>
<th>M</th>
<th>V M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian American</td>
<td>A 4/76</td>
<td>85 95</td>
<td>P</td>
<td>35 95</td>
<td>M</td>
<td>4</td>
<td>40 54</td>
<td>580</td>
<td>630</td>
<td>700</td>
<td>153.1</td>
<td>-54.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 4/76</td>
<td>95 95</td>
<td>P</td>
<td>10 85</td>
<td>F</td>
<td>5</td>
<td>46  73</td>
<td>580</td>
<td>690</td>
<td>74.3</td>
<td>-20.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 4/76</td>
<td>95 95</td>
<td>P</td>
<td>10</td>
<td>95</td>
<td>M</td>
<td>5</td>
<td>40 66</td>
<td>470</td>
<td>760</td>
<td>24.9</td>
<td>-9.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seniors</td>
<td>A 11/76</td>
<td>75 95</td>
<td>NP</td>
<td>35 50</td>
<td>M</td>
<td>5</td>
<td>44 71</td>
<td>500</td>
<td>710</td>
<td>1.4</td>
<td>25.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>95 95</td>
<td>P</td>
<td>23 95</td>
<td>F</td>
<td>5</td>
<td>55  50</td>
<td>610</td>
<td>710</td>
<td>720</td>
<td>21.6</td>
<td>93.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 11/76</td>
<td>85 95</td>
<td>NP</td>
<td>25 70</td>
<td>M</td>
<td>2</td>
<td>30  60</td>
<td>520</td>
<td>550</td>
<td>520</td>
<td>23.6</td>
<td>29.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puerto Rican</td>
<td>A 4/76</td>
<td>85 95</td>
<td>NP</td>
<td>13 30</td>
<td>F</td>
<td>4</td>
<td>29 34</td>
<td>330</td>
<td>400</td>
<td>24.0</td>
<td>16.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>75 75</td>
<td>NP</td>
<td>2</td>
<td>50</td>
<td>M</td>
<td>3</td>
<td>40 38</td>
<td>380</td>
<td>360</td>
<td>20.1</td>
<td>45.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seniors</td>
<td>A 11/76</td>
<td>95 95</td>
<td>NP</td>
<td>10 50</td>
<td>M</td>
<td>3</td>
<td>41 37</td>
<td>440</td>
<td>440</td>
<td>39.9</td>
<td>-43.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 85</td>
<td>NP</td>
<td>17 70</td>
<td>F</td>
<td>3</td>
<td>27  31</td>
<td>300</td>
<td>350</td>
<td>-7.8</td>
<td>-16.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Nonblack</td>
<td>A 4/76</td>
<td>87 85</td>
<td>P</td>
<td>25 95</td>
<td>F</td>
<td>5</td>
<td>48 53</td>
<td>480</td>
<td>490</td>
<td>-6.7</td>
<td>-71.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 4/76</td>
<td>95 95</td>
<td>P</td>
<td>25</td>
<td>95</td>
<td>M</td>
<td>4</td>
<td>41 53</td>
<td>600</td>
<td>630</td>
<td>12.7</td>
<td>36.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seniors</td>
<td>A 11/76</td>
<td>85 85</td>
<td>NP</td>
<td>50</td>
<td>M</td>
<td>4</td>
<td>45 53</td>
<td>540</td>
<td>560</td>
<td>75.2</td>
<td>8.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 85</td>
<td>P</td>
<td>35 70</td>
<td>F</td>
<td>3</td>
<td>42  46</td>
<td>480</td>
<td>550</td>
<td>25.0</td>
<td>62.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 4/76</td>
<td>95 95</td>
<td>P</td>
<td>16 95</td>
<td>M</td>
<td>4</td>
<td>43  53</td>
<td>400</td>
<td>530</td>
<td>30.2</td>
<td>-90.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>95 95</td>
<td>P</td>
<td>35</td>
<td>95</td>
<td>F</td>
<td>3</td>
<td>57 54</td>
<td>570</td>
<td>650</td>
<td>-7.7</td>
<td>53.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seniors</td>
<td>A 4/76</td>
<td>85 85</td>
<td>NP</td>
<td>13</td>
<td>70</td>
<td>M</td>
<td>32 43</td>
<td>430</td>
<td>420</td>
<td>23.5</td>
<td>-32.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 85</td>
<td>P</td>
<td>35 70</td>
<td>M</td>
<td>2</td>
<td>37  33</td>
<td>480</td>
<td>450</td>
<td>88.7</td>
<td>-45.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 4/76</td>
<td>95 95</td>
<td>NP</td>
<td>25 85</td>
<td>F</td>
<td>3</td>
<td>52  50</td>
<td>690</td>
<td>540</td>
<td>165.0</td>
<td>-0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 85</td>
<td>NP</td>
<td>21 85</td>
<td>F</td>
<td>3</td>
<td>44  51</td>
<td>450</td>
<td>510</td>
<td>-19.2</td>
<td>-23.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 4/76</td>
<td>85 75</td>
<td>NP</td>
<td>21 50</td>
<td>M</td>
<td>2</td>
<td>20  24</td>
<td>260</td>
<td>240</td>
<td>31.6</td>
<td>-51.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 75</td>
<td>NP</td>
<td>21 50</td>
<td>M</td>
<td>2</td>
<td>20  24</td>
<td>260</td>
<td>240</td>
<td>31.6</td>
<td>-51.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 4/76</td>
<td>85 75</td>
<td>NP</td>
<td>21 50</td>
<td>M</td>
<td>2</td>
<td>20  24</td>
<td>260</td>
<td>240</td>
<td>31.6</td>
<td>-51.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seniors</td>
<td>A 11/76</td>
<td>85 95</td>
<td>P</td>
<td>25 85</td>
<td>M</td>
<td>5</td>
<td>40 70</td>
<td>400</td>
<td>630</td>
<td>700</td>
<td>-35.3</td>
<td>11.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 85</td>
<td>NP</td>
<td>17 70</td>
<td>M</td>
<td>3</td>
<td>44  58</td>
<td>470</td>
<td>600</td>
<td>79.5</td>
<td>50.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 12/76</td>
<td>85 85</td>
<td>P</td>
<td>70</td>
<td>85</td>
<td>F</td>
<td>3</td>
<td>38 46</td>
<td>420</td>
<td>480</td>
<td>4.3</td>
<td>33.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 11/76</td>
<td>85 95</td>
<td>NP</td>
<td>70</td>
<td>F</td>
<td>4</td>
<td>24  30</td>
<td>230</td>
<td>380</td>
<td>80.5</td>
<td>-2.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 12/76</td>
<td>75 85</td>
<td>NP</td>
<td>13 50</td>
<td>F</td>
<td>2</td>
<td>20  32</td>
<td>260</td>
<td>270</td>
<td>-13.5</td>
<td>12.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*—P = Public, NP = Nonpublic
References


Appendix 3
Rock Report

Disentangling Coaching Effects and Differential Growth in the FTC Commercial Coaching Study

DONALD A. ROCK

The FTC study is a pretest/posttest design where subjects are not randomly assigned to groups. In the nomenclature of Campbell and Stanley (1963), we have a special case of a quasi-experimental design known as the nonequivalent control-group design. In the classic experiment with random assignment the only rival explanation to the treatment being the causal factor is sampling error. However, for the nonequivalent control-group design, both sampling error and possible selection bias are plausible rival hypotheses.

The FTC Boston Regional Office analysis (1978) and the FTC Bureau of Consumer Protection reanalysis (1979) relied heavily on the ability of the traditional analysis of covariance model (ANCOVA) to rule out rival hypotheses. In order to accept the FTC’s interpretation of their ANCOVA results as providing reasonable estimates of the treatment effect (coaching), rather strong assumptions have to be made about the appropriateness and completeness of their covariance adjustment model for the data set. The reanalysis presented in this paper will show in some detail that the verbal data do not support the strong assumptions required to arrive at their point estimates of the verbal coaching effect. The mathematics data will be shown to be more consistent with their specification of the analysis of covariance model and thus more likely to produce reasonable estimates of the math coaching effects. An adjustment procedure appropriate for the verbal data will be outlined and applied to the data. New estimates of the coaching effect will be computed and compared with the results based on the standard ANCOVA model.
The statistical design and data base used in the RRC analysis can best be depicted in the Campbell & Stanley framework as follows:

\[
\begin{array}{l|ccc}
\text{TREATMENT (coaching)} & O_1 & I & O_2 \\
\hline
\text{CONTROL} & O_1 & O_2 \\
& t_1 & t_2 \\
\end{array}
\]

where the O's are observations on the treatment and control subjects at two different points in time, \( t_1 \) and \( t_2 \). The I indicates a formal intervention has taken place for the treatment group at a point in time between the first and second observation. The ANCOVA model traditionally used in estimating the intervention effect in this type of experiment can be formulated in terms of the general linear model

\[
Y_i(t_2) = b_1 Y_i(t_1) + b_2 \phi + e_i ,
\]

where \( Y_i(t_2) \) = posttest score,

\( Y_i(t_1) \) = pretest score,

\( \phi \) = dummy variable \( \{1 = \text{treatment}, 0 = \text{control} \} \) indicating presence or absence of treatment, and

\( e_i \) = random error term that is assumed to be uncorrelated with treatment group membership.

The size of the regression weight \( b_2 \) in relation to its standard error leads to a judgment about the presence or absence of a statistically significant treatment. The sign of \( b_2 \) indicates the direction of the effect. In this particular example, a finding of a significant positive \( b_2 \) would be interpreted as evidence for the presence of a coaching effect. A finding of a negative \( b_2 \) suggests that coaching has a debilitating effect on posttest scores.

A more realistic model might be

\[
Y_i(t_2) = b_1 Y_i(t_1) + b_2 \phi + d' \beta + e_i ,
\]

where \( \beta \) = vector of variables which are assumed to be causes of self-selection, and \( d' \) = vector of regression weights associated with the self-selection causes. Relevant causes of self-selection might be negative self-perception of test-taking skills, motivation, ability to pay for formal coaching, etc.

In the RRC study we have no direct measures of these self-selection variables and thus must rely on proxies. To the extent that the pretest \( Y(t_1) \) and the proxies only incompletely \{or incor-
rectly) measure the causes of self-selection, the estimated regression weight, \( b_2 \), will be biased. In statistical parlance the assumption of the independence of the error \( \epsilon_i \) and the treatment no longer holds. Furthermore, if the coached group has an initially higher mean on the pretest and if the pretest and the other causes of self-selection which are either imperfectly measured or not measured at all are positively related to the treatment effect and the posttest score, then we can expect the ANCOVA model in (2) to overestimate the treatment effect. Conversely, if the missing or incompletely measured control variables are in general negatively related to the treatment effect and posttest score, then one can expect ANCOVA to underestimate the treatment effect. If they are uncorrelated with the treatment and the posttest [as would be the expectation in random assignment], then \( b_2 \) would give an unbiased estimate of the treatment efect. Obviously there are many other more complicated patterns of relationships which could occur and for which no simple predictions can be made about the direction of the bias.

Of particular concern here is the fact that motivation is either incompletely measured or not measured at all by the available demographic control variables. Since one of the consequences of motivation is learning performance, one would expect a measure of motivation to be positively correlated with both performance at the posttest and self-selection for the treatment [coaching]. One might argue that differential motivation would be at least partially captured in the pretest score and/or other covariates in the vector \( z \) and thus be partially controlled, but in general that will not be true if the differential motivation is reflected [as is the likely case] in differential learning or growth rates. That is, even if differential motivation were completely measured, the statistical correction employed by ANCOVA is a static correction. The ANCOVA adjustment only makes a correction for differences found at the time when pretesting takes place. The treatment does not occur at the time of pretesting but at some later date when, if there is differential group growth, the groups will be further apart than they were at pretest time. The traditional ANCOVA correction adjusts for the differences found at pretest time and not the greater group difference that would occur [because of differential group growth rates] just previous to the intervention, which in this case takes place at an unknown period of time after the pretest.

This underadjustment bias in the presence of differential group
growth can be made a little clearer if we examine the ANCOVA model in the following form:

\[ \alpha = [\bar{Y}_p(t_2) - \bar{Y}_c(t_2)] - b [\bar{Y}_p(t_1) - \bar{Y}_c(t_1)] \]

(3)

where \( \bar{Y}_p(t_2) \) and \( \bar{Y}_p(t_1) \) are posttest and pretest means for the coaching program group, \( \bar{Y}_c(t_2) \) and \( \bar{Y}_c(t_1) \) are the parallel measures for the control group, and \( \alpha \) is the estimated treatment effect. The pooled within-group regression coefficient \( b \) is, of course, the adjustment index. If there is reason to believe that the following differential development or growth process is taking place:

![Figure 1](image)

then any adjustment based on the initial group differences at pretest time \( t_1 \) rather than at \( t_x \) will overestimate the treatment effect unless \( b \), the adjustment index, is equal to \( d_2/d_1 \). In fact, any adjustment based on group differences at time \( t_x \) will still overestimate the treatment effect (although substantially less than an adjustment at \( t_1 \)) unless the adjustment index is greater than 1.0.

In general, most pre-post ANCOVAs based on data from nonequivalent control group designs yield \( b \approx 1.0 \). The "net effect" of the adjustment made by the FTC analysts using an ANCOVA model similar to that outlined in equation (2) was equivalent to using a \( b \) of approximately 1.0 in equation (3) for the verbal data and somewhat over 1.0 for the mathematics data. Clearly, the traditional ANCOVA model as posed by the FTC analysts will not always be adequate if individuals are growing at a different rate over time. To deal adequately with data generated by growth processes, corrections for initial differences must take into consideration: (1) the
time variable and (2) measures of differential group growth rate in the absence of formal intervention.

The FTC analysts did not investigate the possibility of differential group growth rates as a rival hypothesis. Their verbal ANCOVA model implicitly assumed (and its validity depended upon it) that the observed group differences at pretest time were invariant over time in the absence of a formal treatment. This assumption is reflected in their use of a "net" correction index of $b \equiv 1.0$. Conversely, since the FTC mathematics ANCOVA yielded a "net" correction index slightly over 1.0, the FTC mathematics results implicitly assumed a growth model. Campbell (1969) suggests that if two groups start out at time 1 with divergent means, those with the higher mean mature or grow at a greater rate than those with the lower mean. Campbell calls this the interaction of selection and maturation. The different groups are members of different populations living in different environments.* The different environments interact with differences in ability and create and maintain different levels of performance and different rates of growth. A special case of this differential growth phenomenon is the so-called "fan spread" model which postulates that increasing variability within group accompanies increasing mean differences over time (Kenny, 1975). The "fan spread" hypothesis is only one possible explanation and is sometimes used to arrive at a correction index when group growth data in the absence of intervention are not available. In its strictest sense a solution is obtained by the use of an overidentifying restriction depending on the stationarity over time of the ratio of mean differences to changes in the within-group variability. Rather than make any such assumptions, we will use the observed data to estimate group growth rates in the absence of intervention.

An interrupted time series design could provide an estimate of the relative growth rate in the absence of a formal intervention. That is, the following design uses growth data gathered between time $t_0$ and $t_1$ to

\[
\begin{array}{cccc}
P & O & O & I & O \\
\hline
C & O & O & O \\
& t_0 & t_1 & t_2
\end{array}
\]

*The FTC analysis of the group demographic variables does provide evidence that individuals in the self-selected population tend to come from different socioeconomic environments than the controls.
yield an estimate of both the program \( P \) and control \( C \) group rate of growth in the absence of a formal intervention. Fortunately, the majority of the individuals who attended the most successful coaching school \( \text{School A} \) had three testings with a coaching intervention taking place between time \( t_1 \) and \( t_2 \). Following the growth model development of Bryk and Weisberg (1977), we assume that the treatment effects a constant increment for each subject exposed to it. Denoting \( G_i(t) \) as the actual growth for individual \( i \) under whatever conditions prevail, we then have at posttest

\[
Y_i(t_2) = G_i(t_2) + \phi_i \alpha,
\]

where \( \phi = 1 \) if the subject is in the coaching program and \( \phi = 0 \) if the subject is in the control group. \( G_i(t_2) \) is the cumulative growth of the \( i \)th subject at the time of posttesting. \( G_i(t) \) is completely determined by two parameters: \( \pi_i \) and \( \tau_i \). That is:

\[
G_i(t) = \pi_i [t - \tau_i]; \ t \geq \tau_i,
\]

where \( \pi_i \) is an individual’s growth rate and \( \tau_i \) is that point in an individual’s development at which he achieves a prespecified although arbitrary level of performance. For the purpose of this study, the \( E(\tau_i) = \tau \), that point in time when the two groups’ growth lines intersect. At this point in time, the groups are at equivalent levels of knowledge. (See Figure 2).

![Figure 2](image.png)

In this study, we are interested in estimating \( \pi \) for individuals in both the coaching program group and the control group based on the time interval between the first and second testing, i.e., before the intervention took place. Thus, the estimation of \( \pi \) for the \( i \)th subject in the coaching program is

\[
\pi_{pi} = [Y_{pi}(t_{i1}) - Y_{pi}(t_{0i})] / (t_{i1} - t_{0i})
\]
That is, the growth rate is the difference between a subject’s test scores at time \(t_0\) and \(t_1\) divided by the time interval measured in months. The expected value for an individual selected at random from the treatment group \(\mu_{sp}\) is \([\bar{Y}_p | t_1] - \bar{Y}_p | t_0] / \bar{T}\), where \(\bar{T}\) is the average time lapse between first and second testings for individuals in the coaching program group. Similarly \(\mu_{sc}\) can be estimated for the control group from the data based on the first and second testing. It should be noted that in the case of random assignment \(\mu_{sp} = \mu_{sc}\).

A more appropriate adjustment index which takes into consideration the possibility of differential group growth rates \(\mu_{sp}\) and \(\mu_{sc}\) as well as different \(\bar{T}\)s can now be derived. In terms of the general linear covariance model, which does not assume differential growth rates, we have the estimated treatment effect \(\hat{a}\) in large samples as

\[
\hat{a} = E[\bar{Y}_p | t_2] - E[\bar{Y}_c | t_2] - b_{lim}[E[\bar{Y}_p | t_1] - E[\bar{Y}_c | t_1]]
\]

(7)

Let \(a\) be an unbiased estimator of the treatment effect under the assumption of a differential group growth model; then

\[
a = [E[\bar{Y}_p | t_2] - \mu_{sp} \bar{T}_{p2}] - E[\bar{Y}_c | t_2] - \mu_{sc} \bar{T}_{c2}]
- b^*[E[\bar{Y}_p | t_1] - \mu_{sp} \bar{T}_{p1}] - E[\bar{Y}_c | t_1] - \mu_{sc} \bar{T}_{c1}]
\]

(8)

where \(\bar{T}_{p2}\) is the average time interval between \(\tau\) and the third testing, \(t_2\), for individuals in the coaching program. Similarly \(\bar{T}_{p1}\) is the average time between \(\tau\) and the second testing for the coached individuals. Parallel notation is used for the control group.

Then \(\hat{a} - a = 0\) when

\[
\hat{b}_{lim} = b^* = \frac{\mu_{sp} \bar{T}_{p2} - \mu_{sc} \bar{T}_{c2}}{\mu_{sp} \bar{T}_{p1} - \mu_{sc} \bar{T}_{c1}}
\]

(9)

Thus \(b^*\) represents the theoretically correct adjustment coefficient if indeed the differential growth rate model holds.

Figure 3 presents means and standard deviations of the verbal test scores at three points in time. Similar data is presented for mathematics test scores in Figure 4. The three testings consisted of the PSAT, the SAT taken for the first time as a junior, and the SAT
taken for the second time as a senior. Only one coaching school (School A) had a large enough subject population to furnish sufficient subjects with all three data points. The coached sample included 192 individuals while the controls numbered 684.

The verbal data in Figure 3 suggests that the observed group differences at posttest may well be reflecting differential growth rates as well as treatment effects. The classic growth phenomenon of increasing means is demonstrated for both the coaching program people and the controls. While there appears to be a significant differential group growth rate in the verbal area, this does
not appear to be the case for mathematics. This seems reasonable since mathematics is a skill which requires practice if one intends to maintain or increase one’s level of achievement.

Additional plots of the total College Board sample who were tested in April of their junior year and retested in November of the senior year are also shown. These plots are presented for 1975 and 1976 populations. Inspection of the verbal plots suggests that gains of 12-17 points from April to November appear to be commonplace for junior to senior retesters. The fact that the control sample shows a gain of 11 points over approximately the same in-
terval suggests that they are reasonably representative of junior-senior retesters. The control group’s rate of gain is consistent over both plotted time periods.

Table 1 presents ANCOVA and growth model adjustment indices and coaching effects, as well as the mean growth rates for the coaching program and control individuals for the verbal and mathematics areas. These estimates of growth are based on the time period between the first and second testing and thus are unconfounded with any formal intervention. Our estimates of growth rates assume linearity of growth but this would seem to be a reasonable assumption given the restricted time period and, even more importantly, given that the observed data for the control group conforms quite closely to a linear growth model. If one were to expect a deviation from linearity, it would more likely be in the direction of a steeper slope (in the absence of intervention) between the second and third testing since the motivational level would be at least as great during this time period as the previous time period.

The estimates of the coaching effect under the growth model are about 17 points on Verbal and 31 points on Math. This finding that the Verbal effect is substantially less than the Math effect is in marked contrast to the ANCOVA estimates of quite comparable V and M effects. The growth model appears to yield an estimate of the Verbal coaching effect that is more consistent with earlier studies and with expectations that Math, being generally more curriculum related than Verbal, might be more responsive to coaching or special preparation.

The Verbal results in Table 1 call into question the ability of the standard ANCOVA to correct for selection effects when they are present in the form of differential growth rates. The ANCOVA adjustment index (b) is approximately 1.0 for the Verbal data and somewhat paradoxically 1.25 for the Math data, which is the one situation which displays little or no evidence of differential group growth. However, since the growth rates appear to be static for the mathematics data, the ANCOVA model, which controls for all available measures of pre-existing differences, would appear to be the more defensible approach for estimating the mathematics "coaching" effect.

An additional point should be discussed here. The original FTC analyses as well as the reanalysis as outlined here do not take into consideration the possible biasing effect of using a fallible covariate (pretest) in adjusting for pre-existing differences. If the reliability of the tests increased with each testing, one would expect
the two groups if truly from different populations to demonstrate increased divergence over time in the absence of intervention. This phenomenon would not seriously affect the growth model approach but very well could have a biasing effect on the ANCOVA estimate of the coaching effect. That is, ANCOVA would yield an overestimate of the coaching effect. Similarly, if the reliabilities are constant but less than 1.0 (as surely they must be), then ANCOVA would underadjust for pre-existing differences on the pre-test, thereby yielding an overestimate of the coaching effect. The data were not available to investigate these possibilities.

### Table 1

Mean Growth Rates, Adjustment Indices, and Estimated Coaching Effects Under Different Model Assumptions

<table>
<thead>
<tr>
<th></th>
<th>Average Growth Rate¹ (Points/Month)</th>
<th>Adjustment Indices and Their Estimates of Coaching Effects²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ*</td>
<td>Growth Model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b₁</td>
</tr>
<tr>
<td>Verbal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>3.294</td>
<td>1.485</td>
</tr>
<tr>
<td>Control</td>
<td>1.654</td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>2.354</td>
<td>1.102</td>
</tr>
<tr>
<td>Control</td>
<td>2.048</td>
<td></td>
</tr>
</tbody>
</table>

¹The average gain in points per month is estimated in the absence of intervention. Growth model estimates b₁ and a₁ are the adjustment index and the estimated effect based on the group growth rate estimates in the first column. The ANCOVA estimates are the standard estimates arrived at using the list of control variables presented in Table 2, Appendix 3. The ANCOVA b₂ is the net adjustment index; when used in the usual ANCOVA equation it yields the effect estimate a₂.

### Table 2

Analysis of Covariance Models

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>Verbal Posttest b</th>
<th>S.E.</th>
<th>Math Posttest b</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT1-V</td>
<td>.8024</td>
<td>.0238</td>
<td>.1538</td>
<td>.0253</td>
</tr>
<tr>
<td>SAT1-M</td>
<td>.1307</td>
<td>.0235</td>
<td>.7760</td>
<td>.0251</td>
</tr>
<tr>
<td>Rank</td>
<td>.1286</td>
<td>.1319</td>
<td>.3201</td>
<td>.1403</td>
</tr>
<tr>
<td>Sex</td>
<td>2.0659</td>
<td>3.5128</td>
<td>13.4507</td>
<td>3.7389</td>
</tr>
<tr>
<td>Latest Grade in English</td>
<td>.5944</td>
<td>.3134</td>
<td>-.5246</td>
<td>.3335</td>
</tr>
<tr>
<td>Latest Grade in Math</td>
<td>-.2336</td>
<td>.2559</td>
<td>.6165</td>
<td>.2723</td>
</tr>
<tr>
<td>High School Type</td>
<td>6.4684</td>
<td>3.6079</td>
<td>-8.5227</td>
<td>3.8402</td>
</tr>
<tr>
<td>Log [Income]</td>
<td>-1.8385</td>
<td>2.8870</td>
<td>-1.0852</td>
<td>3.0728</td>
</tr>
<tr>
<td>Coaching</td>
<td>28.0658</td>
<td>4.0403</td>
<td>27.6247</td>
<td>4.3004</td>
</tr>
<tr>
<td>Years of Math</td>
<td>-2.9292</td>
<td>2.2961</td>
<td>8.7692</td>
<td>2.4400</td>
</tr>
<tr>
<td>Years of English</td>
<td>7.0633</td>
<td>3.0307</td>
<td>-2.8102</td>
<td>3.2258</td>
</tr>
<tr>
<td>Multiple Correlation</td>
<td>R = .8803</td>
<td></td>
<td>R = .8964</td>
<td></td>
</tr>
</tbody>
</table>

133
Summary and Conclusions

The FTC data based on individuals who attended the largest and most successful coaching school were reanalyzed. The examination of testings at three points in time suggested that: (1) the traditional ANCOVA approach used by the FTC was inadequate for the Verbal data because of self-selection effects which were at least partially captured in differential group growth rates, (2) more appropriate growth-related adjustment models yielded Verbal coaching effects about one-half the size reported by the FTC, (3) the Math data appeared to be somewhat more consistent with the ANCOVA model and thus the resulting FTC Math coaching-effect estimate is more likely to be reasonable given the available control variables. This is not to say that the FTC and the present estimates of the mathematics coaching effect are not overestimates (or underestimates), since the only self-selection causes that have been controlled for were those that were reflected in differential growth rates and/or available demographics. One missing piece of information is the reliability of the test scores at each administration. When individuals self-select to treatments, commonly held psychometric wisdom suggests that as the reliability of the pretest becomes significantly less than unity, one can expect the ANCOVA model to underadjust, thereby yielding an overestimate of coaching effect. In general, the extent of this source of error would be small compared to the consequences of using the standard ANCOVA model in an inappropriate situation.

If the reliability of the pretest was different for the two groups, then almost any ANCOVA or variation of ANCOVA would in general yield biased results. It would not be expected, however, that the group reliabilities would be sufficiently different to significantly change the above conclusions. An additional problem with any reliability correction is that while there is evidence here that the groups are from different populations and therefore regressing on different means, there is no way of knowing from the available data what the appropriate means are.
References

Bryk, A. S., & Weisberg, H. I. Use of the nonequivalent control group
design when subjects are growing. *Psychological Bulletin*, 1977, 84,
950-962.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental

Campbell, D. T. Reforms as experiments. *American Psychologist*, 1969,
24, 409-429.

Federal Trade Commission, Boston Regional Office, *Staff memorandom
of the Boston Regional Office of the Federal Trade Commission: The
effects of coaching on standardized admission examinations*. Boston,
Mass.: Federal Trade Commission, Boston Regional Office, September
1978.

Federal Trade Commission, Bureau of Consumer Protection. *Effects of
coaching on standardized admission examinations: Revised statistical
analyses of data gathered by Boston Regional Office of the Federal
Trade Commission*. Washington, DC: Federal Trade Commission, Bu-

Kenny, D. A. A quasi-experimental approach to assessing treatment
effects in the nonequivalent control group design. *Psychological Bul-
letin*, 1975, 82, 345-362.