DOCUMENT RESUME

ED 204 347 TM 007 012

AUTHOR Pike, Lewis W.
TITLE Short-Term Instruction, Testwiseness, and the Scholastic Aptitude Test. A Literature Review with Research Recommendations.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.
PUB DATE 79
NOTE 47p.
AVAILABLE FROM College Board Publication Orders, Box 2815, Princeton, NJ 08541.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *College Entrance Examinations: Guessing (Tests): High Schools: *Instruction: *Program Effectiveness: Research Design: *Research Needs: *Scores: *Test Wiseness: Time Factors (Learning)
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

The research literature on short-term-instruction (STI) and intermediate-term instruction (ITI) for the Scholastic Aptitude Test-mathematical section (SAT-M) and the Scholastic Aptitude Test-verbal sections (SAT-V) was reviewed. Selected studies of STI and ITI for tests other than the SAT-M and SAT-V, and of testwiseness (TW), were included in the survey if they were judged relevant to the question of special instruction for the SAT. The research studies were reviewed and interpreted within the framework of a score components model that posited four content-related and two TW scores components, as well as test-taking confidence and efficiency, that are theoretically subject to STI and ITS effects. In addition, examinee, item, and instructional characteristics were considered as they relate to the score components model. Basic discrepancies between negative and positive findings were noted for both the SAT-M and the SAT-V. There were generally resolved in favor of recognizing meaningful STI effects for the SAT-M, but remain unresolved for the SAT-V. Recommendations were made for SAT-M and SAT-V research allowing STI effects to be partitioned according to examinee, item, and instructional characteristics as they apply to selected test score components. (Author/RL)

College Board
Research and Development Report

Short-Term Instruction, Testwiseness, and the Scholastic Aptitude Test

Short-Term Instruction, Testwiseness, and the Scholastic Aptitude Test

A Literature Review with Research Recommendations

Lewis W. Pike

College Entrance Examination Board, New York, 1979

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Copies of this report may be ordered from:
College Board Publication Orders
Box 2815, Princeton, New Jersey 08541.

The price is $3.

Editorial inquiries about this report should be addressed to:
The College Board
888 Seventh Avenue, New York, New York 10019.

This book, published by the College Board, is based on work commissioned and funded by the Board.

Printed in the United States of America.

Contents

The research on which this report is based was conducted in 1977 by the author. It is being published now to share Lewis W. Pike's conclusions with other interested researchers.

Stephen W. Ivens, The College Board

Abstract

The research literature on short-term instruction (STI) and intermediate-term instruction (ITI) for the SAT-mathematical sections and SAT-verbal sections was reviewed. Selected studies of STI and ITI for tests other than the SAT-M and SAT-V, and of testwiseness (TW), were included in the survey if they were judged relevant to the question of special instruction for the SAT.

The research studies were reviewed and interpreted within the framework of a score components model that posited four content-related and two TW score components, as well as test-taking confidence and efficiency, that are theoretically subject to STI and ITI effects. In addition, examinee, item, and instructional characteristics were considered as they relate to the score components model.

Basic discrepancies between negative and positive findings were noted for both the SAT-M and the SAT-V. There were generally resolved in favor of recognizing meaningful STI effects for the SAT-M, but remain unresolved for the SAT-V. Recommendations were made for SAT-M and SAT-V research allowing STI effects to be partitioned according to examinee, item, and instructional characteristics as they apply to selected test score components.

Introduction

This study was requested by the College Board to provide an up-to-date summary of research findings relevant to the question of special instruction for the Scholastic Aptitude Test. The need for such a review lies both in the continued relevance of the question, and in the fact that the last summary was completed several years ago (College Board, 1968).

Questions regarding special instruction for the SAT remain relevant for several reasons. One reason is that the continued importance of SAT scores to examinees results in a continued pressure to obtain "instruction for the SAT," which in turn leads to an active commercial "coaching" enterprise and to efforts by some public and private schools to provide such instruction. Another is that the changing make-up of the examinee population needs examination. It is entirely plausible, for example, that the more advantaged students represented in most of the studies cited in the College Board booklet, Effects of Coaching on Scholastic Aptitude Test Scores (most of them conducted in the 1950s), were already well prepared to do their best on the SAT to a degree that cannot be assumed for an increasing proportion of the current candidate population, particularly minority and other students outside the mainstream of educational opportunity. Finally, there have been studies of instruction directed either to the SAT or to closely related topics such as the "coachability" of verbal analogies that have appeared since the College Board booklet was published that need to be considered in current thinking and general statements regarding instruction for the SAT.

SCOPE OF THE REPORT

The literature review will cover two interrelated areas of study: (1) studies of short-term instruction (STI) and intermediate-term instruction (ITI) directed specifically toward increasing test scores, with particular emphasis on the SAT-V and SAT-M; and (2) studies of testwiseness (TW) that were not specifically directed to raising test scores. The review of the literature will be followed by recommendations for future research.

Two topics will be considered next that should help clarify the subsequent review of the literature and facilitate the discussion of its implications: the components of observed test scores as they relate to questions of short-term instruction (STI) and testwiseness (TW); and definition of terms.

COMPONENTS OF OBSERVED TEST SCORES

Implicit in many discussions of STI and TW is the assumption that an individual's test score is essentially a composite of the ability or

knowledge for which a person is being tested, testwiseness, "error"-chance factors in the sampling of test items--lucky guesses--and so on. This assumption is often accompanied by the belief that the intended "real" or "true" score on aptitude tests such as the SAT-V and SAT-M is necessarily (by definition) subject only to gradual, long-term change and, as a corollary, a distrust or suspicion of anything that might alter aptitude test scores in a relatively short term (i.e., STI). To put this question in perspective, it is useful to consider the following delineation of the components of observed test scores.

A. "True score" components: e.g., verbal aptitude, mathematical aptitude.
   1. A composite of underlying knowledge (e.g., vocabulary, elementary algebra) and reasoning ability, developed over a long period of time. (Long-term acquisition, long-term retention.)
   2. A state of being well-reviewed, so that the performance to be demonstrated is in line with the individual's underlying developed competence. (Short-term acquisition, short- or medium-term retention.)
   3. Integrative learning, overlearning, consolidation. (Short-term acquisition, long-term retention.)
   4. Learning criterion-relevant, analytic skills (e.g., how to identify the main idea of a paragraph; how to simplify complex quantitative terms before comparing their value). (Short-term acquisition, long-term retention.)

B. Primary test-specific components.
   1. The match between developed ability (including the various score components listed in A above) and test content. Mismatches may occur as gaps in such areas as skill in locating information in reading passages and ability to work with the algebra of inequalities.
   2. General TW--test familiarity, pacing, understanding of general directions, general strategies for using partial information, and so on.
   3. Specific TW--components similar to B2, but in reference to characteristics of specific item formats (such as verbal analogies and quantitative-comparison items), and other item characteristics.

C. Secondary components influencing test taking.
   1. Level of confidence.
   2. Level of efficiency--the ability to use available knowledge and reasoning ability quickly with a relatively low rate of error resulting from working rapidly.

D. "Error." Fluctuations in attention, sampling error, variations in luck when guessing, etc.

SOME DEFINITIONS AND CONCEPTUALIZATIONS

Terms such as STI, ITI, coaching, TW, guessing, and the "aptitude" versus "achievement" distinction are central to discussions regarding special preparation for test taking, and the meanings of these terms tend to vary from one writer to the next. It will be useful, therefore, to give a brief definition of each, as used in this review, and to expand on the conceptualizations where needed.

Short-term instruction (STI). The term STI will refer to attempts to improve test scores by means of a relatively short period of instruction; relatively short, that is, when compared to the amount of time generally considered necessary for any substantial change in the ability or knowledge in question. STI may be directed toward any or all of the components of observed test scores noted above except true-score component A1, which is by definition limited to long-term acquisition. Note that STI for compo-

nents A2, A3, and A4 is in fact directed toward the ability of interest, even though the instruction is short-term. It may be added that in general there is no sharp contrast between education and STI; given appropriate content, STI may properly be viewed as instruction provided in addition to, rather than instead of, conventional long-term learning (i.e., component A1).

Intermediate-term instruction (ITI). As the name suggests, ITI will refer to attempts to improve test scores by means of special instruction for a somewhat longer period than STI but still a short period compared to the amount of time generally considered necessary for substantial changes in the ability in question. Except for the difference in the relative period of instruction, the description given of STI also applies to ITI.

Coaching. This term will refer to a subset of possible STI activities limited essentially to very brief instruction in general testwiseness, such as effective pacing, answering items whenever partial information about them is known, and practice in answering questions similar to those in the target examination. Specifically not included in this definition of coaching is any content instruction beyond that which is merely incidental to the practice sessions. This definition is implicit in the College Board (1968) statement on coaching, in the design of most of the studies reported there, and in the interpretation of their results. It has been fairly widely adopted, as is indicated in a recent statement on coaching made by Anastasi (1976): "Item types on which performance can be appreciably raised by short-term drill or instruction of a narrowly limited nature are not included in the operational forms of the (SAT) tests" (p. 43).

Testwiseness (TW). In essence, TW is a set of skills and knowledge about test taking that enables individuals to display their abilities (e.g., verbal and mathematical aptitude) to their test advantage. A TW component is by no means unique to standardized tests. It is also present in other modes of assessment such as classroom recitation and essay writing.

Early recognition of the TW component in SAT scores is evident from the fact that "From 1926 to 1944 candidates were required to present completed practice booklets before they were allowed to take the test" (Fremer and Chandler, 1971. p. 147). TW instruction is sometimes viewed primarily as an effort to beat the test, with the assumption that testwise examinees will somehow get higher scores than they deserve. For well-made standardized tests, however, clues that offer spurious routes to correct answers are scrupulously avoided, and the opposite, more compelling concern is that examinees who are not testwise may receive inappropriately low scores. Thus, Stanley (1971, p. 364) uses the contrasting term "test-naiveté," and Ebel (1965) notes that "More error in measurement is likely to originate from the students who have had too little, rather than too much, skill in taking tests" (p. 206).

Guessing. Stated simply, guessing consists of answering a test question in the absence of certainty as to the correct response. It may be divided into three categories: guessing that is blind or random, guessing that is spurious or based on a hunch, and guessing based on partial information. In contradistinction to the common feeling that guessing is at least faintly disreputable, the following four points should be noted.

First, guessing is necessary for responding appropriately to the SAT and to most kinds of assessment. Most examinees encounter some test questions about which they have partial information that would enable them to eliminate at least one choice. In such cases they must guess among the remaining alternatives if they are to benefit from their partial information.

Their guessing in such instances benefits not only them but the users of the test scores as well, because only when partial information is used is it possible to give greater credit to those who are partially informed with respect to a given question than to examinees who are uninformed about it.

Second, although not everyone would agree, guessing would appear to be appropriate in situations such as taking the SAT. This point may be clarified by describing contrasting situations. If a student is taking an "open-book" examination, or is writing a term paper, it would indeed be unscholarly and inappropriate to guess or to gloss over points of uncertainty rather than seeking out the needed information. On the other hand, guessing may be inappropriate in a testing situation in which the required information has been clearly specified ahead of time, and mastery of that information emphasized. This would be particularly true if the testing procedures used are consistent with this situation, and guessing on the test is actively discouraged. However, aptitude tests such as the SAT, and even typical large-scale standardized achievement tests, present a test-taking situation that is markedly different. There is not a clearly specified listing of points of information to be mastered, and of course there is no opportunity for seeking additional information as is the case for "open-book" tests. Thus, the test situation, including accompanying directions about guessing, makes it appropriate to guess when answering SAT items.

Third, it may be argued that despite the misgivings of some educators, guessing on tests such as the SAT is not antithetical to good decision making or good scholarship. In most enterprises, whether building bridges or investigating theoretical problems, the point is necessarily reached where information gathering must be terminated and estimations, educated guesses, and the like must be resorted to.

Finally, the net result of guessing on the SAT is fair; over a set of items, partial credit is received for using partial information.

Aptitude versus achievement testing. The literature on STI and TW is sprinkled with allusions to differences between aptitude and achievement tests, generally indicating that STI effects are both more likely and more acceptable for achievement tests than for aptitude tests. Essentially, the distinction is that aptitude tests are more general, more oriented toward reasoning, and less curriculum-bound than are their achievement test counterparts. The distinction becomes problematic when it is then suggested that aptitude tests, by definition, should be relatively impervious to STI. With regard to the components of observed test scores noted above, this need only be true for component A1. Component A2 (effective review) theoretically allows for STI effects on aptitude test scores, because as Carroll (1970) has observed, "The SAT is in truth a test of developed abilities, depending both on general intellective capacities to learn and on an accumulation of knowledge and skills acquired through education in, and experience with, the verbal and mathematical aspects of this nation's culture" (p. 2). STI components B2 and B3 (general and specific TW) apply more potentially to aptitude tests than to achievement tests to the extent that aptitude tests more often resort to more complex item formats such as verbal analogies, data sufficiency items, and quantitative comparisons. There appears to be an increasing tendency toward seeing the distinction between aptitude and achievement testing as one that is relative rather than categorical, particularly with regard to the mathematical area.

Literature Review

Because research regarding the SAT-M is more definitive than that directed to the SAT-V, the two will be reviewed in that order. Selected studies of instruction directed to other aptitude tests and subtests and to achievement tests will be considered. Finally, studies examining selected aspects of TW will be reviewed.

INSTRUCTION FOR THE SAT-MATHEMATICAL

Studies of STI or ITI directed specifically to increasing scores on the SAT-M will be considered in chronological order. Those conducted prior to the Pike and Evans (1972) report will be considered only briefly, because they have been summarized elsewhere (College Board, 1968; Evans and Pike, 1973; Fremer and Chandler, 1971; Pike and Evans, 1972).

The first six of these studies (Dyer, 1953 a,b; French, 1955 a,b; Lass, 1958; French and Dear, 1959; Frankel, 1960 a,b; Whitla, 1962) all involved the use of SAT pretests and posttests. The period of time devoted to STI followed a typical format chosen by the instructors, but generally consisted of group practice with test items similar to those appearing in the SAT-M. All reached the conclusion that score gains attributable to coaching were not sufficient to justify having students invest time in such instruction to improve their scores. In some of the studies of particular subgroups of students and/or particular kinds of items there were instances of meaningful score gains. These instances (as well as any other exceptional finding or observation) will be noted for each of the studies.

Of the last four of the studies directed to increasing the SAT-M scores (Marron, 1965; Roberts and Oppenheim, 1966; Pike and Evans, 1972; McCarthy, 1976), all but the second differ from the first six studies, particularly in that they give emphasis to mathematics content review in addition to other kinds of STI or ITI. The Roberts and Oppenheim study differs from all the others in focusing on students considered to be academically disadvantaged.

Dyer study. In this study, coached students (239 boys) averaged 13 points greater gain on the SAT-M 200 to 800 scale than was observed among the 229 control students in a similar preparatory school. The effect was considerably greater, 29 points, when the comparison was made for students who had taken no mathematics as seniors. The 13-point and the 29-point differences were both statistically significant. Data in the appendix to the Dyer report indicate an average gain of about 15 SAT-M points for the total group of control students.

French study. Here, an overall gain of 18 SAT-M points was observed

comparing coached students' gains in one school to those for control students in two other schools. Boys not currently taking mathematics gained 29 points when compared to one control group, and 9 when compared to another; those taking mathematics gained 19 points and 5 points for the same comparisons. Paradoxically, girls not taking mathematics gained either 5 points or 1 point, whereas those who were taking mathematics showed a coaching effect of 30 points or 20 points. A plausible explanation would be that in both studies the boys currently taking mathematics had a "ceiling effect" on the benefits of review, but that those not taking mathematics were able to get maximum benefit from coaching. For the girls, on the other hand, it may be that those not taking mathematics were victims of what Tobias (1977) describes as "math anxiety," since they appeared to derive no benefit from the brief review that was provided.

Lass study. Comparisons were made of gains between junior and senior year SAT-M scores for students who received no coaching, those who received outside coaching, and those who received a school-provided orientation program. The latter made students familiar with SAT testing procedures and test content but did not involve extensive drill on multiple-choice test questions or other typical coaching activities. SAT-M score gains for the three groups were 53, 64, and 52 points, respectively, from junior- to senior-year test administrations. Thus, there was a slight advantage for receiving coaching. Perhaps more notable are the sizable changes for all three groups compared to the 15- to 20-point gains ordinarily observed over this interval of schooling.

Dear study. This study, reported by French and Dear (1959), was designed to be more intensive than the earlier studies. Classes were much smaller (two students in each), and more time was allotted. However, specifics of content and form of instruction were again left to individual teachers, and the assumption that classes of only two students are optimal is not necessarily true. For students not currently taking mathematics, those receiving coaching gained an average of 28 points more than those not coached. For students who were taking mathematics, the average gain attributed to coaching was only 6 points.

Frankel study. This study involved students at the Bronx High School of Science, which had a record of sending 98 percent of its graduates to college. Nearly all students take four years of mathematics. In this study, coached students received 30 hours of instruction from a commercial coaching school. Those who were coached were reported as experiencing a 9-point loss when compared to the controls. However, Frankel also reported the gain scores for both groups, rather than simply the difference between the two. Control subjects gained 66 points between the May and December or January SAT-M, compared to 57 points for coached subjects. These changes, when compared to an average change in SAT-M scores over a similar interval of 15 points for over 1.6 million students (Pike and Evans 1972, p. 5), suggest that the faculty at Bronx High School of Science were already doing exceptionally well in preparing students for taking the test, whether directly or indirectly. In such a school, there is evidently little need for any additional preparation for test taking.

Whitla study. Like Frankel, Whitla examined the effects of commercially provided coaching for the SAT over a similar time interval between pretest and posttest. Coached students showed no SAT-M gain between the second pretest and the posttest; control subjects gained 6 points. Control subjects were volunteers in the same schools attended by students who had elected to obtain instruction from a proprietary organization.

Marron study. The effects of intensive ITI directed to the SAT were

studied for a total of 714 students in 10 preparatory schools, most of whom planned subsequently to enter one of the United States military service academies. No description was given of the instruction provided, but presumably there was some mix of mathematics content instruction and practice on test items. The overall average SAT-M was changed from 532 to 611, an increase of 79 points.

Roberts and Oppenheim study. This study was undertaken to determine whether coaching for the SAT could effectively raise scores of students from academically disadvantaged backgrounds. Volunteers in eight Tennessee high schools were randomly assigned to an instructional group for SAT-M or to a control group. Mean PSAT pretest scores were equivalent to about 330 on the SAT-M scale. The instruction had essentially no effect. In discussing the results the authors noted the limited amount of instruction (15 half-hour sessions) and indications of inadequate motivation. The results may offer another instance of "math anxiety," where the gap between student preparation and the level of the test was too wide to be tolerated easily.

Pike and Evans study. When this study was undertaken, the SAT-M was made up of two kinds of items: regular mathematics (RM), which is conventional and straightforward, and data sufficiency (DS), which is much more complex. At issue was whether to replace these entirely or in part by a third kind of item, quantitative comparison (QC). The question was motivated by a desire to reduce the SAT-M testing time, and by the fact that QC items had been found to be remarkably efficient. There was concern, however, that the QC items might be relatively more susceptible to STI than the RM and DS items already in the SAT-M. The study was designed to examine the relative susceptibility of each of the three kinds of mathematics aptitude test items to STI. In the process, evidence directly related to instruction for the existing SAT-M was also obtained.

A separate instructional program was developed for each of the three kinds of test item. Male and female high school junior volunteers in each of 12 schools were given a pretest consisting of a complete SAT, a supplementary test made up of QC items, and tests parallel to these as posttests several weeks later. In the intervening period experimental subjects received seven three-hour instructional sessions directed to one of the three formats. Control subjects were provided the same instruction after the posttest.

The instruction differed from that reported in previous studies in that it was highly systematic, and control of instruction was provided through students' workbooks and teachers' lesson plans. It was more comprehensive than in the studies reviewed above in respect to the components of observed test scores noted earlier. Included was systematic content review (component A2), involving overlearning or mastery learning of basic geometric principles, computing averages, etc. (A3); learning such analytic skills as simplifying quantitative terms that are being compared (A4); filling in informational gaps such as computations using inequalities (B1); and teaching both general and specific aspects of TW (B2 and B3).

Mean gains of nearly a full standard deviation obtained by subjects instructed for the complex item types (DS and QC), compared to gains of about one-fourth standard deviation by control subjects on these tests, were of statistical and practical significance. The RM items were interestingly less susceptible to STI. Coached students gained about one-half standard deviation compared to a gain of one-fifth standard deviation by control subjects. Although none of the subjects were instructed for all three kinds of SAT-M items, changes in SAT-M score suggest the STI effects that could be ex-

pected if similar methods and materials were used and adapted to cover both kinds of items then in the SAT-M. The RM subjects gained 43 points on the SAT-M, compared to an 18-point gain by controls. Thus, the STI effect of RM instruction was 25 points, even though just two-thirds of the SAT-M items were of that type. A judicious combination of RM and DS instruction, kept within the 21 hours of instruction, would be expected to yield an STI effect of about 33 points.

It is reasonable to ask whether gai , demonstrated immediately following STI will decrease soon thereafter. Of the 377 experimental subjects who took the December posttest, 288 subsequently took the regularly administered SAT the following April. There was an average gain of an additional 24 SAT-M points.

McCarthy study. The mathematics department at Longmeadow High School, Longmeadow, Massachusetts, provides a course "to help students review the mathematics course that they have learned in their high school career. . . . in preparation for the fall administration of the SATs." The instruction includes several practice tests but also provides a formal review of mathematical concepts. SAT-M score gains have been recorded for participants and nonparticipants in the program, comparing spring junior-year scores to winter senior-year scores. For winter 1974-75, mean gains for instructed students (N = 30) and for a random sample (N = 50) of noninstructed students were 38 and 21 respectively. For winter 1975-76, mean gains for 60 instructed and 60 noninstructed students were 57 and 16 respectively.

INSTRUCTION FOR THE SAT-VERBAL

Studies of STI directed specifically to increasing scores on the SAT-V will also be considered in chronological order. In most instances, these will be the same studies for which SAT-M effects were considered. Five of the first six of these studies have been summarized in the 1968 College Board booklet on the effects of coaching.

Dyer study. In this study, coached students averaged only 5 SAT-V points greater gain on the SAT-V 200 to 800 scale than was observed for control students in a similar school. Differences were also examined by item type (analogies, sentence completion, antonyms, and reading comprehension). Those for analogies were significant at the .05 level; the size of the difference was not reported. Data in the appendix to this study show a mean gain of about 30 SAT-V points for the control students.

French study. SAT-V instruction was provided in two schools, and students in a third school served as controls without instruction. In the first school, gains attributed to coaching were 18 points, in the second school, 5 points. In the former, instruction on analogies accounted for two-thirds of the total effect of coaching, and in the latter the effect was due almost entirely to antonyms. These differences underscore the variability in instruction, but they also suggest possible differential susceptibility to instruction, depending on item type; analogies are the most likely to yield a sizable effect from STI.

In school A (no instruction), school B (verbal coaching), and school C (verbal and mathematics coaching), SAT-V gains were 28, 33, and 46 points respectively.

Lass study. Comparisons were made between junior and senior SAT-V scores for students who received no coaching, those who received outside coaching, and those who received a school-provided orientation program. SAT-V score gains for the three groups were 41, 44, and 53 points. It is interesting to note here that the orientation program seems to have been more beneficial

than the coaching. Once again, all three groups showed sizable average gains.

Dear study. This study (French and Dear, 1959) showed essentially no effects due to SAT-V instruction for coached students.

Frankel study. Among students at Bronx High School of Science, uncoached students gained 38 SAT-V points compared to 47 points for those who received commercial instruction, a difference of only 9 points. Gains for both groups were larger than typically observed for May-December/January score changes (18 points on the average for SAT-V), which again suggests an accelerated rate of growth at this school. The SAT-V gains were not as pronounced as for the SAT-M, on which controls and coached students gained 66 and 57 points respectively.

Pallone study. Pallone (1961) reports the effects of two programs of instruction for the SAT-V, one STI and the other ITI. He did not attempt instruction for the SAT-M. Pallone deliberately designed instruction to go beyond the "coaching" that has so regularly been found ineffectual and focused instead on the reading, vocabulary, and logical reasoning abilities that the SAT-V is assumed to measure. The STI was in the form of a very systematic study program involving instruction in intensive reading skills, skimming, critical reading, reading comprehension exercises, and the analysis of verbal analogies and was provided in daily 90-minute sessions over a six-week period. Thus, it seemed most directed to score component A4 (learning criterion-relevant analytic skills) and also covered component A3 (integrative learning), B1 (filling in gaps in developed ability), and B3 (TW specific to analogies).

The 20 participating students showed an average gain of 98 SAT-V points. Because there were no control subjects there is no direct way to subtract from this the effects of practice and growth in order to estimate the STI effect. Using the gains experienced by controls at the Bronx High School of Science as a rough (and probably conservatively high) estimation of control subject gains, the effects of STI in the Pallone study would be estimated at approximately 60 points.

The ITI program involved daily 50-minute instructional periods over a five-month interval. Program content was similar to the STI except for a substantially greater amount of instruction. About 80 students completed the ITI program, and for these the average SAT-V score gain was 109 points. The 20 students receiving STI also received the ITI, and there was an overall score gain for these students of 122 points.

Whitla study. Students receiving commercial instruction for the SAT-V gained 11 points more than the control subjects between pretest and posttest. (Controls gained 20 points between the two testings, and 39 points altogether between the pre-pretest junior-year SAT and the posttest taken as seniors.)

Marron study. Following intensive ITI in the 10 preparatory schools, the average SAT-V score changed from 471 to 528, a gain of 57 points.

Roberts and Oppenheim study. Volunteers in six Tennessee high schools were randomly assigned to a PSAT-V instructional group or to a control group. Mean PSAT-V pretest scores were equivalent to about 315 on the SAT-V scale. As with PSAT-M instruction, programmed instruction was provided in 15 half-hour sessions. Instructed students gained the equivalent of 7 SAT-V points, and controls lost 7 points. The control group's loss of points was apparently due to motivational problems.

Coffman and Neun study (1966). This study was undertaken to determine the effect of a presumably typical accelerated reading course on SAT-V scores. Three groups of college freshmen took part in the study, each

receiving 45 to 50 hours of instruction as part of a college-credit course emphasizing speed with relative accuracy. There were no control subjects. Mean score changes were +4, +10, and -29. The last change is statistically significant, suggesting that instruction for that group may actually have hindered effective performance on the SAT-V. The authors described the results as being in disagreement with Pallone's findings. However, since the instruction appears to lack most of the features provided by Pallone for increasing verbal reasoning powers, the two studies seem scarcely comparable.

INSTRUCTION FOR TESTS OTHER THAN THE SAT

Two studies (Marron, 1965; Jacobs, 1966) involving instruction for the College Board English Composition Test (ECT) are relevant to the question of instruction for the SAT-V. It may be noted, for example, that two of the four item formats used in the SAT-V (reading comprehension, and antonyms) could as well be viewed as testing the attainment of reading skills and vocabulary respectively. Furthermore, the ECT contains complicated item formats, and, as a result, instruction directed in part to the relevant TW components may have implications for TW instructions for other relatively complex formats such as verbal analogies in the SAT-V and data sufficiency or quantitative comparison items in the SAT-M.

Two additional studies (Moore, 1971; Whitely and Dawis, 1974) are addressed specifically to questions regarding instruction for answering analogy items.

Marron study. Of the students taking SAT pretests and posttests in the Marron study, 347 also took the ECT on both occasions. The average gain on the ECT 200 to 800 scale was 83 points, from a pretest score mean of 458.

Jacobs study. Student volunteers in each of six schools were randomly assigned to a group receiving instruction or a control group. The STI consisted of six three-hour sessions. About nine hours were spent directly on criterion skills (score components A2, 3 and 4), and about nine hours on specific TW (score component B3) related to item format. In each school, specific TW was directed to two of the three ECT item formats (sentence correction, construction shift, and paragraph organization). The ECT was administered only after the experimental subjects had received instruction. In two of the schools, involving a total of 36 students receiving STI and 44 control students, there were only negligible differences between scores for the two groups. In the other four schools, involving 91 instructed and 87 control students, mean differences ranged from 44 ECT points in one of the schools to 73 points in another. Such clear evidence of STI effects occurring in some schools but not in others suggests that the specifics of STI provided by different instructors may have a marked effect on the outcome of an STI experiment.

Moore study. Instruction for answering verbal analogy items was provided to graduate students by a booklet directed to two aspects of the task: understanding the format of the question, and learning to recognize specific classes of relationship. The 38 subjects were randomly assigned to an experimental or a control group. A 75-item analogy test with a somewhat more cumbersome format than that used for the SAT-V was subsequently administered. Students receiving STI averaged 44.3 items correct compared to 39.7 for controls, a difference of about three-fourths of a standard deviation. The number of subjects was very small, so these results should be considered tentative. If the findings replicated, however, they would demonstrate that even brief instruction to relatively sophisti-

cated examinees can make a difference in performance on verbal analogy items.

Whitely and Dawis Study. The subjects were 184 students randomly selected from the class lists of two inner-city high schools in St. Paul, Minnesota. Those selected were randomly assigned to one of five treatment groups or to a control group. Verbal analogy items used for the study had an unusually low vocabulary level, so that answering the items would depend primarily on the ability to educe relationships rather than on word knowledge. The pretest and posttest each consisted of a 41-item analogy test. Fifty analogy items were used for all five treatments. One treatment involved practice on the 50 items without feedback, and another involved practice with feedback of the correct answer. The other three treatment groups also had practice with the 50 items, with instruction interspersed between item subsets that was addressed primarily to helping students learn to recognize such categories of relationships as "opposites," "class membership," and "functional." The three groups receiving instruction differed in that one was instructed under the condition of feedback and structural aids (in which 10 additional analogies were presented with structural labels and arrows indicating the related pair), another with feedback only, and the third with structural aids only. It was found that the only experimental group to perform significantly better than the controls was the one receiving instruction combined with both feedback and the diagrammatic structural aid. All six groups had pretest means of about 24 and standard deviations of about 9. The control group gained about 2.3 items correct, the "instruction plus feedback plus structure" group gained about 6.3, and the other groups between 3.4 and 4.0 items correct.

These results indicated that well-designed STI (only 50 minutes were used for the intervention) can sometimes meaningfully increase performance on analogy items, and that practice, beyond that obtained in taking the pretest, even with feedback, had no meaningful effect unless it was supplemented by carefully designed instructional materials.

STUDIES EXAMINING TW

The topic of TW is frequently investigated in studies not involving instruction for specific tests or subtests. Because of the importance of TW as a component of test scores, and the implications of this component regarding test validity and fairness, some of the general findings in these studies of TW will be reviewed here. These will be clustered in several categories. First will be studies or commentary relevant to adequately defining TW (Alker, Carlson and Hermann, 1967; Crehan, Koehler, and Slakter, 1974; Diamond and Evans, 1972; Ebel, 1965; Millman, Bishop, and Ebel, 1965; Stanley, 1971). Second will be the topic of guessing (Cronbach, 1970; Diamond and Evans, 1973; Flaugher and Pike, 1970; Lord, 1964; Lord, 1975; Slakter, 1968 a,b; Pike and Evans, 1972; Pike and Flaugher, 1970; Thorndike, 1971). Third is the related topic of risk taking (Slakter, 1967; Slakter, 1969; Slakter, Crehan, and Koehler, 1975; Swineford and Miller, 1953). Fourth is another topic related to guessing, that of answer changing (Bath, 1967; Jacobs, 1972; Lynch and Smith, 1975; Mueller and Schwedel, 1975; Mueller and Wasser, 1977). The fifth topic is TW related to particular kinds of items. These include studies of verbal analogies (Connolly and Wantman, 1964; Gentile, 1966; Gentile, 1968; Gentile, Kessler, and Gentile, 1969; Willner, 1964), and of reading comprehension items (Pyrczak, 1974; Vernon, 1962).

On defining TW. Earlier in this paper, TW was defined as "that set of

skills and knowledge about how to take a particular test that allows individuals to display their abilities to their best advantage." It will be useful at this point to consider other definitions, explicit or implicit, commonly used in the testing literature when discussing TW. The definition most often encountered in the literature is that proposed by Millman, Bishop, and Ebel (1965): "'Test-wiseness' is . . . . a subject's capacity to utilize the characteristics and formats of the test and/or the test-taking situation to receive a high score. Test-wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures" (p. 707).

Implicit in both definitions is the possibility that some aspects of TW are necessary if examinees are to receive proper credit for the knowledge or ability being tested, and that other aspects of TW may allow examinees to receive more credit than is their due, i.e., the test-sophisticate may be able to "beat the test." The Millman et al. definition appears to elicit the latter concern. Typical of reformulations of their definition is that used by Diamond and Evans (1972); who define TW as ". . . the ability to respond advantageously to multiple-choice items containing extraneous clues and to obtain credit on these items without knowledge of the subject matter" (p. 145). Another instance of picking up on the beating-the-test aspect of the Millman et al. definition is found in Alker et al. (1967) who state: "Defined in this way (Millman et al.), testwiseness emphasizes the use of the format of the test rather than its content to achieve a higher score. . . ." (p. 11). Note that they could as well have said "in addition to" instead of "rather than." On the other hand, awareness of a need for the opposite concern, particularly with respect to well-constructed objective tests, is evident in statements by writers such as Ebel and Stanley, as was noted earlier. With regard to tests such as the SAT, a concern that examinees should have the required TW to cope well with the test as a vehicle through which they are to demonstrate their verbal or mathematical ability would appear to be more compelling. This is in keeping with the recommendations of Crehan et al. (1974) who, upon demonstrating that some examinees are consistently low on TW across tests, noted that TW can never be fully eliminated as a component of standardized tests and suggested that ". . . perhaps more thought should be given to the teaching of tw to students low in tw" (p. 211).

Studies of guessing. A central part of TW is knowing when and how to guess, where guessing is defined as answering a test question in the absence of certainty as to the correct response. The problem is especially troublesome for objective tests of ability, in part because multiple-choice questions heighten our awareness of the guessing component, and in part because a general test of ability, especially if it is of an appropriate difficulty level for a given examinee, will have many items for which the examinee is neither certain of the correct answer (and therefore has no need to guess) nor so totally uninformed as to be reduced to blind guessing. Although blind or random guessing is the kind that comes to mind initially and is discussed most often, it is probably the least likely to occur. Most guessing decisions will involve choosing whether to answer a question or not, when the basis for doing so is either partial information or a spurious hunch or feeling.

Much of the research literature on the question of guessing on objective tests is focused on the use or nonuse of a "correction formula" for guessing. Diamond and Evans summarized this literature in 1973 and found little basis for any conclusive answers. When not certain of the answer to a question examinees vary considerably in their willingness to guess, even when

there is no penalty for doing so. To reduce the score differences resulting from this factor, the standard guessing penalty is often imposed. Ironically, as Slakter (1968a) noted, research has shown that even when there is a penalty for guessing most examinees would do better if they guessed more. He elaborates the point in another article (1968b) where he notes that the scoring penalty directions tend to influence most those students who are already reluctant to guess, resulting in the guessing penalty becoming a penalty for not guessing. Responding to the argument that examinees often fail to use the best guessing strategy but instead omit items on which they could do better than chance (and thus could expect to benefit even when there is a penalty), Lord (1975) comments that "Perhaps the difficulty can be corrected by giving better test instructions. If not, it may be time for children in school to be taught how to behave effectively when taking a test" (p. 8).

Because of the problems such as those noted above associated with the standard correction formula, many testers and educators advocate "rights only" scoring, accompanied by instructions to answer every item. At least two problems emerge from this approach. First, many examinees still decline to answer every question. Second, as noted by Lord (1964) and others, "Forced random guessing necessarily increases the error of measurement present in the test scores" (p. 746). Yet another problem is that requiring examinees to answer all questions shifts the problem from knowing when to guess to knowing how to guess. This is exemplified by findings of Flaugher and Pike (1970) who were examining relative amounts of random-like answering behavior by inner-city students on the Preliminary Scholastic Aptitude Test (PSAT), a test that was very difficult for them. Using an index developed by Pike (Pike and Flaugher, 1970), substantially more random-like behavior was found in the error responses of the inner-city students than in those of the norming population for three item formats: reading comprehension, sentence completion, and antonyms. For analogies, however, inner-city error responses were as systematic as those for the norming population. However, actual performance on analogies was comparatively worse than that on the other item formats, suggesting that the inner-city examinees were being systematically attracted to error choices for this item type. In subsequent work with the Graduate Record Examinations (GRE), Pike (1978, unpublished) found that low-scoring examinees scored at worse than chance level on more than half the analogy items, but only rarely on other types of verbal aptitude items. This occurred in spite of the option of omitting and a clearly stated penalty for guessing. It is likely that forcing such examinees to answer all analogy items would serve to decrease their scores.

Studies of risk taking. Examinees having a given level of information or confidence for answering a test question may differ considerably in the tendency to guess despite any directions to the contrary. These differences are commonly referred to as differences in risk-taking (RT) behavior. Attempts to reduce this variability in deciding when to guess have been discussed in the immediately preceding paragraphs as a part of the general question of guessing. Studies directed to risk-taking behavior will be considered next.

Swineford and Miller (1953) investigated TW by administering a vocabulary test that included items having nonwords and words extremely unlikely to be known for which answering would strongly imply guessing. The test was given under three different instructions: (1) encouraged to guess; (2) told not to guess; and (3) not instructed about guessing. It was found that (1) there was some guessing in all three sets of directions; (2) there was a slight difference between no instructions about guessing and instruction

favoring guessing (it appeared easier to inhibit guessing than to encourage it); and (3) there was little relationship between guessing, risk taking, and ability.

Various measures of risk taking have been compared by Slakter (1967). In a later study (1969) he pointed out that examinees who do not take risks tend to be penalized on test scores, and he also noted that this tendency usually generalizes across different tests. Still more recently Slakter, Crehan, and Koehler (1975) reported on a longitudinal study of RT tendency. They found again that RT was relatively stable across tests for an individual examinee at a given time, but found longitudinal changes that point to the fact that RT tends to decrease over grades 5 to 9 and then becomes relatively stable, at least through grade 11. They noted an important implication of this finding, that the contribution of RT strategy toward maximizing test scores actually tends to become less between grades 5 and 9.

Studies of answer changing. Yet another aspect of the question of guessing, which is more in the realm of how to guess than when, is that of answer-changing behavior on multiple-choice tests. The topic is of interest because student opinion and much of the advice given by educators runs directly counter to most research findings. Two excellent summaries on the question are provided by Lynch and Smith (1975) and Mueller and Wasser (1977). Among the more recent studies of interest are those of Bath (1967), Jacobs (1972), and Mueller and Schwedel (1975). The following conclusions emerge from these studies.

1. Most examinees express the belief that it does not pay to change answers.

2. Most examinees do change answers but typically on only about 4 percent of the questions.

3. In fact it generally does pay to change answers. Typical findings are that there are about two favorable changes for every unfavorable change.

4. Gains drop off as items get relatively more difficult.

5. Higher scoring examinees tend to benefit more from changing answers than do those who score lower.

Studies of testwiseness (TW) for specific item types. The "how" of effective guessing becomes particularly central when attention is given to specific kinds of items, especially those that are relatively complex. In surveying studies of TW, studies directed specifically to reading comprehension items and to verbal analogies, both of which are relatively complex, were found particularly relevant.

Vernon (1962) examined the assessment of reading comprehension of British and American examinees by comparing free-response data from essays, fill-in sentences, etc. to multiple-choice responses. He found a test-sophistication factor in the multiple-choice responses of British examinees who were generally unfamiliar with such tests that was much less evident in American responses. The difference was more pronounced for the reading comprehension items than for the more straightforward vocabulary questions. Pyrczak investigated an intriguing aspect of testwiseness by studying the effects of answering test items for reading comprehension independently of the accompanying passage. In one study (1972) he found that examinees rely on various sources of information and misinformation when answering such questions in the absence of the reading passages and also make use of interrelationships among the items in a given set. In a subsequent study (1974) he reduced these sources of answering strategy and found that examinees were still able to perform at a better-than-chance level, presumably by such devices as selecting statements of

general principles rather than specific facts, and by selecting the most general of several principles presented.

Willner (1964), working with analogy items drawn from a wide variety of tests (Miller Analogies Test, Army Alpha, Otis Beta, etc.), found that about half the items could be answered correctly on the basis of word association alone, i.e., without having to educe relationships for a given item and then solving the analogy on the basis of the educed relationships. He recommended that analogy items that are substantially free of the word-association effects on analogy solving be constructed and used in tests. He added the impressionistic observation that in some instances word associations led to the wrong answer, and that some examinees who might have solved an analogy on the relational basis appeared instead to have been distracted from doing so by the strong associational attraction of one of the error choices.

Connolly and Wantman (1964) used "think aloud" data elicited from nine subjects in solving verbal analogies to observe analogy-solving processes. The observations were largely impressionistic. Two impressions were relevant to the present review. First, the words provided in the alternative answer choices influenced how the stem words were interpreted. The subjects were often observed to revise the relationships they had established for the stem pair of words to fit the demands of the first option. Second, it was observed that the subjects seemed to differ considerably in their methods of attacking or analyzing test items. Both observations are relevant to score component B3, "specific TW," and the second has implications regarding component A4, "relevant analytic skills."

Gentile (1966) and Gentile, Kessler, and Gentile (1969) have also examined performance in solving verbal analogies (drawn from recired SAT-V items), giving primary consideration to the amount of score variance attributable to word associations. In the 1969 study "associative relatedness" was found to account for 28 to 50 percent of the score variance. Their discussion suggests that they consider the effect of associative relatedness to be an inherent part of analogy items, a position that contrasts with Willner's discussion in which the availability of an associational basis for answering analogies without resort to educing relationships is viewed as a problem that can be remedied by changes in test construction. Gentile (1968) also examined the effect of sociocultural level and the knowledge of definitions on analogy solving. The latter was done by observing the effect of providing definitions of words appearing in the analogies. He found the effects both singly and in combination to be weak.

Summary and Interpretation of Findings

Some general considerations will be noted first that provide a useful framework for doing so. Following that, findings relevant to the SAT-M and those having a bearing on the SAT-V will be considered, with results derived from studies of tests other than the SAT cited where appropriate. Findings from studies of TW will be summarized last.

GENERAL CONSIDERATIONS

Design characteristics. The first consideration is that of the basic design features of the studies themselves. The studies differ substantially in such important variables as the use or nonuse of control groups, the selection of control groups (ranging from using groups of students in schools generally comparable to those the experimental subjects are in, to the use of random assignment), the number of subjects, and the use either of pretest and posttest data or of alternatives to that procedure.

Summarizing mixed findings. The next consideration is the question of how research findings should best be interpreted, particularly when making comparisons across studies. In principle, a single study showing substantial positive gains cannot be countered or refuted by any number of studies failing to get positive results. The only near exception would occur in the event of a well-designed replication study that failed to show similarly positive results. In that case, there would be a discrepancy needing further study and resolution. Similarly, it would be fallacious to infer, from mixed results across studies on a topic such as STI effects, that across-study inconsistencies justify the conclusion that there are no meaningful effects. As exemplified in Jacobs' (1966) discussion of differences on English Composition Test score changes from one experimental group to another, mixed results can mean that an effort should be made to find out why instruction was effective in some places but not in others. This observation is particularly true when making comparisons between studies in which little account was taken of either examinee or instructional characteristics. A third observation is that there has been a considerable emphasis in most discussions of STI on the overall magnitude of its effects, with little consideration given to differences among examinees, STI curriculums, or item formats and other item characteristics, especially when stating final conclusions.

The tendency to consider only overall average results of STI, together with a polarization of attitudes toward STI as being essentially good or bad, has tended to distract attention from analyses and interpretations that could lead to a more cumulative, orderly base of information regard-

ing the kinds of STI that are effective, for whom, and in what particular areas of test-preparedness (i.e., to what test score components). In the summaries for the SAT-M and SAT-V, two distinctions will be used. Examinee, instructional, and item characteristics will constitute one distinction. Within each of these, the several kinds of test score components will be the other distinction.

The STI, ITI, regular-instruction continuum. The next two considerations are interrelated. First is the realization that STI, ITI, and "regular instruction" are not categorically distinct but rather fall along instructional continua, differing in the relative amount of time required to produce a gain and in the relative directness to which the instruction is oriented toward a specific test. The related consideration is the recognition that score gains associated with various score components may be influenced by many sources. These sources may be either self-study or guided study directed to relevant context critical analytic skills, TW, etc., the use of regular classroom and textbook materials, commercial coaching materials, the SAT descriptive materials, and so on.

The limits of STI and ITI effects. A final consideration in reviewing and interpreting the STI and ITI literature is recognition of the theoretical and empirical limits of their effects. A lack of awareness of these limits is evidenced in statements such as that made by Paul Houts, who is described by Downey (1977) as a testing expert with the National Association of Elementary School Principals. Houts, in reaction to a decision to teach test-taking strategies in the Washington, D. C., public school system said: "Unless someone intercedes at some point, this is going to go on and on. What happens when everyone is coached and everyone does well on the tests?" (Downey, 1977, p. 30). He evidently had tests such as the SAT in mind, as he next described circumstances ". . . under which the universities would have to stop requiring them." A more realistic level of concern but one still describing a potentially excessive STI effect that is quite unlikely to occur, given the theoretical limits of such effects, is exemplified in the latter half of the following statement by Coffman and Neun (1966): "If special preparation and coaching provide long-term improvement in the student's vocabulary, reading speed and comprehension, then they serve a useful purpose for the student. If, however, they lead to an inflation of the student's test score without improving the underlying ability, the student may simply gain admission to a college where his probability of doing successful work is low" (p. 1).

The score components model. These concerns about unlimited or excessive STI effects may profitably be addressed from the standpoint of the score components model. Component A-1 (general level of developed ability) is presumably by far the largest component of an examinee's SAT-M or SAT-V score and is by definition subject to meaningful gains only through long-term acquisition. Effects of STI addressed to component A-2 (review) are subject to two limiting factors. First, "review" presupposes that relevant material had already been learned at an earlier time. Second, the effects for a given examinee are necessarily limited by the extent of need for review. For the SAT-M the need for review is likely to be substantial for some examinees, but for the SAT-V this need is likely to be rather more limited, as virtually all examinees spend much of their time involved in the use of spoken and written English. Before leaving consideration of the review component (A-2), it should be noted that even instances of rather large score gains due to effective review for, say, the SAT-M could not properly be described as instances of excessive gain. In such cases an examinee's performance has been raised to a level consistent with his or

her underlying developed competence. We see, then, that it is the nature rather than the amount of STI gains that determines whether they may be properly considered as excessive.

Score component A-3 (integrative learning, overlearning) again presupposes prior learning at some reasonable level of mastery. Component A-4 (knowledge of criterion-relevant, analytic skills) is at least conceptually subject to STI effects, and if this component were shown to be subject to STI it would in no way invalidate the test. Evidence that such abilities are subject to STI should give us more comfort than discomfort, although it would heighten our awareness of possible disparities in the quality of education, whether short term or long term in acquisition. There is little hard data on the topic. It was addressed most directly, perhaps, in the work of Bloom and Broder (1950) and less directly with regard to the SAT-M by Pike and Evans (1972) and for the SAT-V by Pallone (1961). The paucity of data suggesting STI effects for component A-4 may suggest that meaningful gains in this realm, even given excellent instruction for teaching these analytic skills, are likely to be observed only for students who had developed a "readiness" for such gains. A student who reads widely and with avid interest but has not honed his or her analytic reading skills may be such a person.

The next three score components to consider with regard to limits of STI effects are those specific to the activity of test taking itself. Component B-1 (the match between the domain of the examinee's developed ability and test content) is likely to yield only slight STI effects if examinees are clearly aware of the test content domain, and if the test does not contain an undue number of items requiring basic knowledge most of them do not have. For example, STI for solving inequalities is more likely to have a meaningfully large effect on SAT-M scores to the extent that (1) the test has many items calling for this ability, (2) many students have not routinely learned this ability; and (3) many examinees are unaware of the fact that such items are included in the SAT-M.

Score component B-2 (general TW; test familiarity, appropriate pacing, understanding general directions, knowing when and how to guess, etc.) is susceptible to STI effects almost entirely to the extent that the examinee is initially test-naive. Thus, for the most part, any score increase due to STI directed to component B-2 is evidence of having helped students receive the credit to which they are due, rather than having fostered any kind of "beating the test" resulting in "excessive" score gains. Note that gain of this sort is an increase (rather than an inflation) ". . . of the students' test score without improving the underlying ability" that need not imply that "the student may simply gain admission to a college where his probability of doing successful work is low" (Coffman and Neun, 1966, p. 1).

The next score component, B-3 (specific TW; similar to general TW but referring to item format and other item characteristics), is the only component that poses a problem regarding possible "excessive" gain from STI. The problem arises in the case of complex item formats which, in their complexity, tap a kind of methods variance conceptually independent of the mathematical or verbal aptitude the SAT is intended to measure. Vernon (1954), in reviewing the British literature on coaching, concluded that more complex item formats are likely to be more coachable. Loret (1960), in his review of SAT content from the test's inception in 1926 to 1960, made the same observation, and noted a steady trend in both the mathematical and verbal parts of the test toward simpler, more straightforward item formats. Nevertheless, for pragmatic reasons there remain in

use item formats that are sufficiently complex to allow an undesirably large STI effect favoring students who are given help in learning how to deal with the item format complexities. Aside from dropping such item formats altogether, the problem can be reduced in the following ways: (1) by keeping the number of such items proportionally low; (2) by imposing appropriate test specifications within item format (c.f., Willner's suggestion, noted earlier, for minimizing the role played by word association in solving verbal analogies); (3) by expanding and clarifying directions given within each test; and (4) by providing wide dissemination of information describing these item formats and instruction about how to cope with their complexities. To the extent that these four measures are taken, the magnitude of STI effects for component B-3 will tend to fall within acceptable limits.

Although the final pair of score components, C-1 (level of confidence), and C-2 (level of efficiency), are in large measure spin-offs of the preceding seven, STI may include direct attempts to ensure that these benefits do indeed follow from instruction directed to the other score components. Here again it may be noted that even instances of large score gains resulting from changes in the score components in question are instances of helping examinees to receive appropriately higher scores, rather than helping them make excessive gains that might be both unfair and a disservice to the examinees by making their scores unrealistically high.

FINDINGS REGARDING THE SAT-M

A basic discrepancy. In summarizing the findings of studies of STI or ITI for the mathematical sections of the SAT we begin with a basic discrepancy. The overall conclusions of Dyer, French, Dear, Lass, Frankel, Whitla, and Roberts and Oppenheim are essentially negative, whereas those of Pike and Evans, McCarthy, and Marron are positive. The former studies show overall average score changes attributable to STI ranging from slight losses to gains up to about 20 SAT-M scale points. Some of these differences were statistically significant, but none were considered meaningfully large. By contrast, overall STI effects in the Pike and Evans study are gains conservatively estimated at about 33 SAT-M points, and those in McCarthy's 1975-76 data at about 41 points. Overall instructional effects averaged by Marron over 10 preparatory schools yielded a gain of about 79 SAT-M points. Among the other studies in this review, average control group gains ranged from 15 points in Dyer's study to 66 in Frankel's, and there was a median gain of 31 points. Using the latter as a rough estimate of what control subjects might have gained in the Marron study, the effect of ITI would be estimated as 79 minus 31 equals 48 points.

Interpreting the discrepancy. In considering the discrepancy between studies in which positive overall results were reported and those in which negative results were reported, we may consider how to interpret the discrepancy, how seriously to take the positive results, and then examine why the discrepancy was observed. In interpreting the discrepancy, it should be recalled that in principle even a single study showing substantial effects cannot be refuted by any number of studies failing to do so. It follows, of course, that mixed results across studies cannot be dismissed as simply indicating that meaningful effects were somehow due to happenstance, or that mixed results indicate basically no effect over the set of studies summarized.

Credibility of the positive findings. On the other hand it is reasonable to demand of a study that obtains positive results contrary to most other

studies on the same topic that its design, and the results reported, lend strong support for the conclusions before they are fully accepted. In the case of the Pike and Evans study, some 500 students were involved. The STI effects were found to be both statistically and meaningfully significant for all 24 experimental groups, distributed over 24 different teachers in 12 different schools. Random assignment of students to either the coached or control groups was a strong design feature, as was random assignment of experimental and control students to either an A, B or a B, A sequence of pretest and posttest. The latter feature ruled out possible score equating problems that were invoked as possible reasons for some of the larger gains observed in other studies. The McCarthy study was less formal. The report compared the score gains of 60 students who had taken an SAT-M preparatory course for school credit to score changes of 60 students in the same school who did not take the course. The Marron study involved 714 students in 10 schools, most of whom were in preparatory programs designed to help them increase their SAT and College Board Achievement Test scores prior to entry in United States military service academies. There were no control subjects.

Explaining the discrepancy. We turn next to some observations of why the differences in overall conclusions may have occurred. In all the studies for which negative conclusions were reached, the STI tended to be brief and relatively uncontrolled, emphasizing test-taking practice and offering at most only incidental instruction directed to mathematics content. Also included was some instruction for general TW. In most instances the students in these studies were attending prestigious private or public schools where students are generally well test-prepared; i.e., well-reviewed, practiced, informed on questions of general and specific TW, etc. The major exception was the Roberts and Oppenheim study, in which it is likely that the opposite situation interfered with possible STI effects; the students were apparently just too educationally disadvantaged to expect to achieve meaningful gains from short-term instruction directed to the PSAT. Another consideration is that most of the studies reporting negative overall conclusions had serious design problems, particularly in the lack of adequate control groups. In some instances there were no control groups, and in others the experimental and control subjects were in separate schools, and as a result, STI effects were confounded with school effects.

In contrast to the description of the studies yielding negative overall findings, the three studies providing positive overall results may be characterized as having generally involved more hours of instruction. The instruction was moderately to highly controlled and deliberately included appropriate mathematical content. The participating students in these studies tended to be at neither of the extremes of test-preparedness or test-sophistication noted for the other group of studies. Thus they seemed to have a need for STI or ITI and a readiness to benefit from it.

A final consideration in the discrepancy between the two groups of studies is the possibility that they demonstrated clear-cut STI or ITI score gains by means of extreme interventions, rendering them not truly generalizable to the usual concerns and questions about coaching or STI/ITI effects. The generalizability of the Pike and Evans study is supported by the following observations: (1) the instruction was short term, involving only 21 hours of classroom participation; (2) it was effective in all 24 classes despite differences in schools, teachers, and groups of students; and (3) all 500 plus students involved were volunteers who had planned to take the SAT. The instruction was effective over the full range of their initial SAT-M scores of 250 to 650. It should be noted that a central

premise in undertaking the study was that well-developed instructional materials would be essential to a successful program of STI. Student workbooks and teachers' class outlines were developed to facilitate instruction that would systematically incorporate effective mathematics content review (A-2), overlearning/mastery-learning of basic facts and mathematical operations (A-3), extensive guided practice in responding systematically and analytically to individual test items (A-4), filling in informational gaps such as computations involving inequalities (B-1), and teaching general and specific aspects of TW (B-2 and B-3).

The generalizability of the Marron findings is supported by the observations that substantial average SAT-M increases were observed in all 10 participating schools, and the curriculums in all instances were whatever teachers in those schools decided upon. The main question to be raised about whether extreme interventions would limit the generalizability of the Marron results is that of the total amount of time devoted to instruction, which in this case consisted of a full semester devoted explicitly to raising test scores. On the one hand, this large amount of time spent on instruction clearly sets the Marron data apart from those of all the other studies considered in this review. On the other hand, the question of whether long periods of instruction would make meaningful differences on SAT scores is one that has regularly been given serious consideration. Carroll (1970) suggests that "In the case of students who have had average educational experiences but who make low scores on the SAT, one may estimate that even a year of rather intensive remedial instruction would not generally suffice to make a dramatic improvement in test performance . . . ." (p. 4). The College Board Commission on Tests (1970) suggests that ". . . students cannot late in high school hope to improve their performance on the SAT appreciably by studying for it . . .," and then adds: "This is not to say that the abilities that are measured on the Scholastic Aptitude Test are impervious to change; it is to say that if verbal and mathematical aptitude; especially verbal aptitude, can be developed within the length of, say, a school year, no one has yet demonstrated a way to do it" (Vol. 1, p. 12). What seems to be called for is an increased awareness of an underlying continuum for STI, ITI, and regular instruction in which the total amount of time, as well as the relative amount, is a factor. The natural limits on the amount of instruction that might be profitably devoted to general and specific TW, and even to content review, are such that any extended instruction designed to increase test scores will necessarily devote a large proportion of time to regular instruction. It is instructive to note that although average SAT-M gains between the junior year (usually April or May) and senior year (usually December or January) tests are usually between 12 and 20 points, control group gains in the studies reviewed here were often substantially larger. Those for two schools in the French study were 31 and 42; those reported by Whitla, Lass, and Frankel were 31, 53, and 66 respectively.

The McCarthy findings provide a sampling of what might be expected from well-run school classes designed specifically for preparing students for the SAT-M. As such they would generalize to that part of the coaching or STI question. That is, is the goal of increasing SAT-M scores by a well-run course one that is unattainable? For that matter, considering the kinds of instruction most likely to produce meaningful gains (effective review, integrative learning, the teaching of analytic problem-solving skills, instruction to overcome test-naiveté, etc.), is such a goal necessarily one that is corrupt and unworthy? It would seem that judgments of such programs should be suspended until one has considered the specifics of what is

actually done, and that the limits of possible score gains linked to the several score components should also be considered.

Applying the score components model. Having considered the general outcomes of instruction for the SAT-M, and the questions that arose from a disparity between those studies failing to show an STI or ITI effect and those succeeding in doing so, we may next consider the findings not as overall outcomes but as outcomes related to the several score components.

Consider first the four score components having to do with developed mathematical aptitude. It will be recalled that the three components subject to STI effects (A-2, 3, and 4) were generally present in the studies that demonstrated STI or ITI effects and absent in those that failed to do so. We have also noted that limitations in component A-1 (developed ability) may have contributed to the lack of STI effects in the Roberts and Oppenheim study where the gap to be bridged may have been simply too large for STI to have an effect. On the other hand, the Pike and Evans instruction was effective over a wide range of initial SAT-M scores. The Frankel data showing control subject gains of 66 points provide an instance of substantial growth in component A-1. This is an interesting discovery not only because of its magnitude but also because for most students this is apparently the effect of studying advanced levels of high school mathematics (most students in the school take four years of mathematics). These findings suggest that although the mathematics required to answer SAT-M items is intentionally limited to ninth- or tenth-grade content, mathematics beyond that level serves not only as review but also to facilitate answering SAT-M items. This in turn suggests that for mathematics the aptitude-achievement distinction is relative and implies as well that one way to increase mathematical aptitude as measured by the SAT-M is to take additional courses in that subject area.

The importance of component A-2 (review) is supported by data in three of the studies that reported no meaningful overall STI effects. The studies by Dyer, French, and Dear all showed STI gains of 28 or 29 SAT-M points for examinees not currently studying mathematics but much smaller gains for those who were taking mathematics courses. Some support for the possibility that instruction for components A-3 (integrative learning) and A-4 (analytic skills) may lead to a subsequent increased rate of growth in mathematical reasoning ability is provided in the Pike and Evans study, where it was observed that participants not only gained between pretest and posttest but gained an average of 24 additional points between the posttest and the post-posttest that was taken four months later.

We may next examine STI or ITI effects related to the three score components that have to do directly with test taking. This instruction is a kind of "teaching to the test", but as noted earlier its impact is to help students overcome test-specific obstacles that cause them to receive inappropriately low test scores. Component B-1 (the match between an examinee's developed ability and the test content domain) was addressed as part of the content review in the Pike and Evans study, and presumably in those of Marron and McCarthy as well. It would be desirable to use diagnostic test information as well as item content information in those and in future studies to see whether filling specific gaps such as computing averages and solving inequalities has a demonstrable effect.

All studies presumably gave at least some attention to component B-2 (general TW). If, however, there is any strong conclusion to be reached from the studies reporting no meaningful STI effects, it is that instruction for general TW in the form of a few general rubrics such as "use your time well," "answer if you think you know the correct choice or if you can

eliminate at least one alternative," and of loosely structured group practice and discussion sessions is quite consistently ineffectual. This of course has direct implications regarding the probable value of much of the commercially provided test coaching. The aspect of general TW given particular attention in Pike and Evans was that of knowing when and how to guess, given partial information. It was found that there was considerable confusion on the part of students and teachers alike on questions of the scholastic propriety, fairness, and efficacy of guessing when partially informed, and a related confusion regarding the implications of the formula score that "corrects" for guessing by subtracting a fraction of a point for wrong answers. Classroom demonstrations of the results of guessing when there was no information, and again when either two or three of five choices could be eliminated, allowed students in each class to derive the conclusion that over a set of items "partial credit is given for partial information." This component of TW should also be examined for its effect on test-taking behavior and on test scores. Component B-3 (specific TW), particularly for the relatively complex item formats (data sufficiency and quantitative comparison), was also given considerable attention in the Pike and Evans study, which was probably responsible for the greater STI effects observed for these formats than were found for the much simpler "regular mathematics" item format.

Confidence and efficiency (components C-1 and 2) in test taking are most likely to increase if substantial efforts on the earlier score components have been made. Thus, in the three studies involving content instruction, it is very likely that at least some gains attributable to the secondary effects of increased confidence and efficiency in test taking also occurred. To enhance this effect, Pike and Evans incorporated occasional timed practice tests that were tailored to the instruction previously received, in order to provide the students an awareness of having increased their test-taking capabilities.

FINDINGS REGARDING THE SAT-V

Would finding STI effects be feasible? It is a common observation that verbal aptitude is not likely to be as subject to coaching or STI effects as is true of mathematical aptitude. In the preface to the Pike and Evans (1972) monograph, for example, Kendrick stated that: "By now it has been fairly definitely settled that the verbal part of the Board's Scholastic Aptitude Test (SAT) is impervious to coaching. The mathematical part seems similarly, though perhaps not so thoroughly, proof against special preparation, but the question of mathematics is complicated by the fact that some students do not take mathematics in their senior year of secondary school, and lead lives very nearly undisturbed by quantitative thought. For them, it is only reasonable that a little review or warming-up would be helpful. . . ." (p. v). We have noted earlier (page 25) the College Board Commission on Tests' statement that concludes, ". . . if verbal and mathematical aptitude, especially verbal aptitude, can be developed within the length of, say, a school year, no one has yet demonstrated a way to do it" (emphasis added).

The above considerations, and the important role mathematics content instruction appeared to have in the studies showing meaningful STI gains for the SAT-M, make any study purporting to show major SAT-V gains appear suspect. However, a comparison of SAT-M and SAT-V findings among seven studies reporting coaching/STI effects on both (Dyer, French, Lass, Dear, Frankel, Whitla, and Roberts and Oppenheim) can serve to check on this

general impression. For the seven studies, high, median, and low STI effects for the SAT-M were 18, 6, and -9 points; those for the SAT-V were 18, 11, and 0. The surprising result, then, is that overall the SAT-V gains were strongly equivalent to those observed for the SAT-M. Having made this comparison another can readily be added: the gains made by control subjects in these seven studies. High, median, and low score changes by control subjects on the SAT-M were 66, 23, and 0; those on the SAT-V were 41, 34, and -7. Here, too, the comparability is more than might have been expected. The Marron study also provided data for both SAT-M and SAT-V (for ITI) but did not include control subjects in the study design. In that study, gains on the SAT-M and SAT-V were 79 and 57 points respectively. The difference is very likely attributable to the fact that most of the students in the Marron study majored in engineering or related fields, and were also preparing for the College Board Mathematics Achievement Tests. Consistent with this conjecture is the fact that the largest control group gain on the SAT-M, 66 points, was observed for students at the Bronx High School of Science, nearly all of whom were taking four years of high school mathematics.

Having made these SAT-M and SAT-V comparisons, and in the process having learned that STI effects for the SAT-V would not be nearly so aberrant as might have been expected, we may proceed to a summary of the SAT-V findings themselves.

Another basic discrepancy. In summarizing the findings of studies of instruction directed to the SAT-V we again note a basic discrepancy. The overall conclusions of Dyer, French, Lass, Dear, Frankel, Whitla, Roberts and Oppenheim, and Coffman and Neun are essentially negative, whereas those of Pallone and Marron are positive. Furthermore, the findings of four additional studies that are relevant to the SAT-V but not directed specifically to that test also reach positive conclusions regarding score gains attributable to STI or ITI. These are the Marron and Jacobs studies of instruction for the English Composition Test (ECT), and the Moore and the Whitely and Dawis studies of instruction for answering verbal analogy items.

The first group of studies showed overall mean SAT-V gains ranging from -9 to 18 points. Unfortunately neither Pallone nor Marron had control subjects to allow an estimation of score gains specifically attributable to the instruction program. However, the gains in both studies were impressively large. For 20 students receiving STI, and for 80 receiving ITI, Pallone reported mean SAT-V gains of 98 and 122 points respectively. For the 700 students in 10 schools who received ITI, Marron reported an average SAT-V gain of 57 points. In the Marron ITI study, about 350 students also took the English Composition Test (ECT) pretests and posttests and showed an average gain of 83 points on the 200-800 ECT scale. Again there were no control subjects. Using random assignment of volunteers to either instructed or control groups, Jacobs investigated the effects of 18 hours of STI on ECT scores in six schools. Only a posttest was given. In two schools the difference between coached and control students was negligible. In the other four schools involving about 90 instructed and 90 control subjects, mean differences ranged from 44 points in one school to 73 in another.

Moore's study of instruction for analogies would suggest that even very brief instruction can make a meaningful difference in performance on verbal analogies for highly able students. The differences he observed would translate to a gain of about two or three analogies on the SAT-V, which would result in an SAT-V gain of perhaps 10 or 15 points. Although

this is not a large gain on the overall SAT-V score, it would be a meaningful effect as a component of that score, limited to the analogies part of the test. The Whitely and Dawis study showed gains quite consistent in degree to those reported by Moore, the difference being that the group studied were inner-city high school students rather than graduate students.

Interpreting the discrepancy. Confronted again with mixed results across studies we note once more the logical primacy of studies demonstrating effects over those failing to do so, but also reiterate that if they are to be fully accepted, positive STI conclusions must be strongly supported by research design and data. In this respect there are shortcomings in both studies addressed directly to raising SAT-V scores. The Pallone STI findings are based on only 20 experimental subjects, a number small enough to indicate clearly the need for replication before great confidence can be placed in the findings. Furthermore, both the STI and the ITI effects were observed in a single school, and the lack of control subjects leaves open the question of how much of the observed gain was attributable to the programs of special instruction and how much to other factors operating in the school in question. The fact remains, however, that the gains were extraordinary. Control subject gains on the SAT-V in the superior schools studied by Lass, Frankel, and Whitla were 41, 38, and 39 points respectively. If we then estimate that school effects and any other sources of growth and practice in the Pallone school would ordinarily be about 40 points, the average gains attributable to STI and ITI (involving 80 students) would be 58 and 82 points. Using the same estimation of expected control subject gains, Marron's data would indicate SAT-V gains of about 17 points. Thus the students appeared to make gains only slightly greater than they could have expected from attending an exceptionally good high school over the same period of time. Even though the average gains on the SAT-V for students taking the SAT in April or May of one year and again in December or January of the next are usually in the neighborhood of 15 to 25 points, the 57 SAT-V point gain observed for the Marron study is large enough at least to raise doubts about the Commission on Tests' statement about raising verbal aptitude scores within a year, particularly since 10 different schools were involved.

Jacobs' finding of gains ranging from 44 to 77 points on the ECT are not only substantial, particularly as they were obtained with only 18 hours of instruction, but are also impressive in the sense that the research design was strong, with random assignment of subjects to control or experimental groups. The question is whether these findings on an achievement test can be interpreted as relevant to the SAT-V, an aptitude test, particularly since achievement tests with their content orientation are generally considered to be more susceptible to STI. Arguing for the relevance of Jacobs' findings are three observations: (1) achievement tests such as the ECT are viewed as becoming increasingly more like aptitude tests as efforts are made to have questions that will generalize across many school curriculums; (2) parts of the SAT-V, particularly antonyms and reading comprehension items, are measures of vocabulary and reading ability that could as well be viewed as achievement measures; and (3) the ECT contains complex item formats, and results of instruction for coping with these complexities may have implications for possible vulnerability of complex SAT-V item formats (particularly analogies) to similar kinds of instruction.

Marron's finding of an 83-point gain on the ECT for some 350 students serves primarily as a rough confirmation of Jacobs' findings, although ITI was required to do it. Moore's data must be considered as tentative, in part because there were only 19 experimental and 19 control subjects, and

in part because the item format used was rather more cumbersome than that employed in the SAT-V. The Whitely and Dawis study involved 184 students from two high schools, giving an adequate data base from which to work, and in addition it involved a rather sophisticated experimental design. The major question about the generalizability of their data to the SAT-V is that the researchers went to considerable lengths to keep all the analogy items in the study at an unusually low vocabulary level. Although the vocabulary load is also kept reasonably low in analogies used in the SAT-V, many of the more difficult items involve fairly difficult words in order to test the ability to recognize subtle relationships. It may well be that the Whitely and Dawis study is directly relevant to possible score changes for very low-scoring subjects on the analogies part of the test, but this would have to be established in further studies.

Explaining the discrepancy. The next question is why the differences in overall conclusions may have occurred. In most of the studies for which negative conclusions regarding STI were reached, the instruction tended to be brief, relatively uncontrolled, and not directed toward verbal abilities, although an emphasis was placed on individual or group practice in test taking. The Coffman and Neun study departed somewhat from this pattern in that it was designed to determine the effect of a presumably typical accelerated reading course on SAT-V scores. This course involved about 50 hours of instruction as part of a college-credit course emphasizing rapid reading with relative accuracy. The Pallone STI study was comparable to the first seven negative studies in the number of hours spent on instruction; his ITI study was comparable to the number of hours of instruction in the Coffman and Neun investigation. The difference in results appear not to lie in the number of hours of instruction. The Pallone instructions for both STI and ITI differed sharply from those given in any of the negative studies, in that Pallone's instruction was deliberately designed to go beyond the "coaching" that had so regularly been found ineffectual. Instead, the instruction focused directly on reading, vocabularly, and verbal reasoning abilities that the SAT-V is intended to measure. The program was highly systematic and controlled, involving instruction in intensive reading, skimming, critical reading, exercises in answering reading comprehension items, and solving verbal analogies. Marron's study was characterized by the large amount of time involved (a full semester directed expressly to raising selected test scores,, although the amount of time devoted to preparation for the SAT-V is not clear. In any event, some kinds of verbal content instruction can be assumed and perhaps instruction directed to specific item formats as well. The studies of instruction for analogy solving (Moore; Whitely and Dawis) are not necessarily inconsistent with results in the studies reporting no meaningful overall gains on the SAT-V. This will be given further comment.

As was true for the Pike and Evans study of SAT-M instruction, the Pallone study of STI and ITI for the SAT-V differed most markedly from the others yielding negative results in the degree to which instruction was substantive and controlled, with emphasis given to effective review (A-2), integrative learning (A-3), the teaching of relevant analytic skills (A-4), and instruction specific to item format characteristics (B-3). On the one hand, this suggests that the generalizability of the Pallone results is limited to STI or ITI efforts that have a similarly strong content orientation, and perhaps specific TW instruction as well. On the other hand, these characteristics of the Pallone instruction clearly fall within the sphere of STI and ITI questions raised in various College Board and other statements on these topics, and by student, parent, and

professional education organizations. Generalization from the Marron results for SAT-V instruction is limited by the recognition that a considerable amount of instructional time was required to obtain the gains reported.

It is interesting to compare the importance of instructional content and the amount of instruction as they affect SAT-V scores. This is most evident in comparing the Pallone study to that of Coffman and Neun. The considerable amount of time spent in developing reading skills in the Coffman and Neun study yielded trivial gains and even losses in SAT-V scores, whereas the sharply focused curriculums used in STI and ITI in the Pallone study yielded sizable score gains. This difference between comparatively passive, unfocused study and active study directed to specific skills runs counter to the common feeling represented by French and Dear's (1959) conclusion that, rather than seeking coaching, an eager College Board candidate " . . . would probably gain at least as much by some review of mathematics on his own and by the reading of a few good books" (p. 329).

Item format differences. Only one of the 10 studies of SAT-V instruction reported differences by item format. This may have been in part because not many items of any one kind were present, since four item types were used, thus making comparisons risky, and in part because in most if not all the studies attention was focused on the overall results. This is unfortunate, because there is good reason to believe that because of differences in format complexity some item types may be more susceptible to instruction than others. In the French study, SAT-V instruction was provided in only two of the three schools. In the first school, two-thirds of the 18-point gain attributed to instruction was observed for analogies. For the second school, in which a 5-point gain was observed, nearly all the effect was due to antonyms. The difference between the two schools is perhaps best attributed to differences in instruction, the latter not having been closely monitored or controlled. In any event, the analogies effect noted in the one school is consistent with general evidence regarding the relationship between STI effects and item complexity, and with the studies of Moore and of Whitely and Dawis that were directed specifically to verbal analogies.

FINDINGS REGARDING TW

Defining TW. Again, TW will be defined as the set of skills and knowledge about how to take a particular test that allows the individual to display his or her abilities to the best advantage. Implicit in the definition is the recognition that some aspects of TW must be used if the examinee is to receive proper credit for the knowledge or ability being tested, but that other aspects of TW, such as taking advantage of "specific determiners," may allow the examinee to receive more credit than is appropriate. The latter aspect, however, is likely to be at a bare minimum for professionally developed tests such as the SAT.

Guessing. It was noted above that guessing, which may be defined as answering a test question in the absence of certainty as to the correct response, usually involves either a more or less spurious hunch or feeling, or the use of partial information, and is seldom the sort of blind selection that often first comes to mind when the term is used. It was also noted that partial information situations in which guessing is an appropriate behavior are necessarily a part of most objective testing, particularly when the test is at an appropriate level of difficulty.

Considerable thought and research have been given to the question of whether to use a "correction formula" to compensate for individual differ-

ences in guessing tendencies. The results are far from conclusive. Arguments for and against the use of correction formulas were also given earlier. The main conclusions to be drawn from these are that: (1) more information is needed on the subject to resolve differences in findings and conclusions; (2) better within-test or before-test answering instructions may be needed (Lord, 1975); and (3) both "rights only" and "correction formula" scoring procedures pose answering dilemmas to examinees, with the former emphasizing the decision of how to select an answer and the latter emphasizing that of whether to select an answer when in doubt.

Risk taking (RT). Individual differences in guessing tendency at a given level of uncertainty of the correct answer and under a given set of instructions about guessing may be described as differences in risk-taking (RT) behavior. One set of basic findings reported above regarding RT was that of Swineford and Miller (1953), who studied RT under instructions that encouraged, discouraged, or were neutral to guessing. They found that (1) there was some guessing under all three sets of directions, (2) instructions inhibiting guessing were more effective than those encouraging it, and (3) there was little relationship between RT (deciding when to guess), and ability. Another basic finding was that of Crehan, Koehler, and Slakter (1974), who found that an individual's RT tendency is relatively stable across different tests at a given time, but that RT tends to decrease over grades 5 to 9, then becomes relatively stable, at least through grade 11. They noted the implication of this finding, that the contribution of RT strategy toward maximizing test scores actually tends to become less between grades 5 and 9.

Answer changing. The question of whether to change test answers moves from the question of when to guess, toward that of how to do so. Excellent summaries of studies of answer changing are found in Lynch and Smith (1975), and in Mueller and Wasser (1977). Some of the conclusions generally agreed upon are listed on page 18.

TW for reading comprehension items. In considering TW as it applies specifically to particular item types, the shift from when to guess to how to guess is particularly evident. In a comparison of free-response and multiple-choice testing of the reading comprehension of British and American examinees, Vernon (1962) found a test-sophistication factor in the multiple-choice responses of British examinees, who were generally unfamiliar with such tests, that was much less evident in American responses. The difference was more pronounced for the relatively complex reading comprehension items than for the more straightforward vocabulary questions. This would suggest a relatively greater need for TW instruction for students on the more complex item format, reading comprehension. Two strategies were observed by Pyrczak (1972, 1974) in studies of answering behavior when the reading passages were not available. One made use of interrelationships among the items in a given set that accompanies a given reading passage, and another used such devices as selecting general principles rather than specific facts.

TW for verbal analogies. Connolly and Wantman (1964) used "think aloud" procedures with nine subjects and provided an impressionistic report of analogy-solving processes. One conclusion was that words among the alternative choices influenced how the stem words were interpreted. Another was that the students differed considerably in their methods of solving the analogy problems. These observations suggest the need for instruction directed to score components A-4 (relevant analytic skills) and B-3 (specific TW).

Other studies have examined the relationship between word associations and the solving of verbal analogies. Willner (1964) demonstrated that on many verbal analogies (drawn from a variety of tests other than the SAT); nearly half the items could be answered correctly using word associations alone, i.e., without having first to educe the relationships for a given item and then solve the analogy on the basis of the educed relationships. He noted that in some instances word associations tended to hinder rather than facilitate solving particular analogies, and thus the opposite effect seemed to have occurred. His proposed solution to the problem is to construct analogy items that are substantially free of the word association effects. This seems clearly desirable, since the use of facilitative word associations to get a higher score will give some students an unfair advantage; and the susceptibility to the distracting power of other word associations will put test-naive students at a disadvantage. Even if the two effects were well balanced across a set of items, the problem remains that some meaningful part of score variance will occur because of this factor, rather than to examinees' relative ability to solve verbal analogies, i.e., to educe and subsequently use structured relationships between pairs of words. Another way of reducing the problem is to provide instruction in solving analogies. It may be that simply expanding the within-test directions to include one sample item and its solution would be adequate.

Recommendations for Future Research

Specific recommendations for research on short- and intermediate-term instruction for the SAT, testwiseness, and related topics will be preceded by a discussion of the objectives toward which the research would be directed and a discussion of general research design considerations derived from an evaluation of the studies reviewed in this survey.

RESEARCH OBJECTIVES

The immediate objectives of the research to be recommended will be presented after discussing the ultimate objectives toward which these would be directed.

Ultimate objectives. There are three ultimate objectives toward which the research would be directed. The first is to maximize the fairness and validity of the SAT with regard to its short-term and intermediate-term instruction (STI and ITI) score components. The second is not to discourage concern and activity regarding test-preparedness, but rather to foster realistic understanding and expectations regarding possible outcomes of STI and ITI. The third, which would derive from the pursuit of the first two, is the emergence of a more basic understanding of the processes involved in test taking and contributing to aptitude test scores.

In considering these objectives, the score components model will again serve as the organizing principle. Differences in component A-1, aptitudes that have developed over a long period of time, do not fall within the purview of this survey, because the question of special instruction for the SAT, whether as STI or ITI, is by definition excluded from consideration for that component. The final component in the mode (D-1, error variance) is also excluded by definition, since "error" as used here in its traditional psychometric sense is score variance not attributable to the factors being considered. The remaining eight components are all subject to various STI and ITI effects and as such are those with which we will be concerned.

The issue of test fairness, to which the first research objective is addressed, is necessarily raised if there are meaningful STI and ITI score effects because of differences in the availability of instruction and even in the awareness of its possible effects. The fairness of the SAT with respect to the effects of special instruction can be maximized in four ways. The first is by informing examinees and educators of STI components that may increase academic aptitude performance (as distinct from underlying academic competence). These instructional components correspond to score components A-2 (review), A-3 (integrative learning, overlearning), and A-4 (learning relevant analytic skills). The second way is by minimizing the

occurrences of mismatches between the basic skills and knowledge required for demonstrating scholastic aptitude that are assumed in constructing the SAT, and the actual distribution of these skills and knowledge in the examinee population. This requires the discovery of such mismatches and the subsequent change of test content specifications and test directions, as well as the provision of within-test information (such as defining the notations of inequalities) and the dissemination of a clearly phrased statement about the skills and knowledge that are assumed in constructing the test. These activities correspond to score component B-1.

The third way of maximizing the fairness of the SAT is by minimizing the testwiseness (TW) score components, B-2 and B-3. This can be done by limiting the need for TW and by making it as generally available as possible. Limiting the need for TW can be accomplished primarily through test content decisions. As noted earlier, many of the changes in the SAT from 1926 until recently have been in the direction of reducing item format complexity. A direct result of such changes is a reduction in required TW. Other test characteristics such as the distribution of item difficulties, the presence of word association effects in verbal analogies (whether hindering or facilitative), the clarity and completeness of within-test directions, and the degree of speededness all influence the amount and kinds of TW required by the test. The TW that is needed can be made available primarily through the use of widely disseminated test familiarization materials.

The final means of ensuring maximum fairness of the SAT with respect to special instruction is that of helping students find ways to develop test-taking confidence and efficiency. The first three means of ensuring fairness are probably necessary to bring about desirable score changes attributed to these last components, C-1 and C2, but are not always sufficient.

The steps just described for maximizing the fairness of the SAT would also tend to increase test validity. STI designed to increase scholastic aptitude performance (e.g., components A-2, 3, and 4) helps ensure that the full potential aptitude developed over a long period of time is at a state of readiness to be demonstrated on the SAT. To put it another way, the individual's scholastic performance that is shown when taking the SAT can be brought in line with his or her full underlying competence, and it is the latter that the SAT is intended to measure and is considered to be the best index of academic potential. Similarly, if there are specific gaps in the individual's requisite knowledge for coping with the test content that can be taught in a short time (B-1), then the score he or she would receive after such STI is more representative of the individual's level of developed aptitude than the score he or she would have received without such instruction. The effect on test validity of minimizing the TW score components (B-2 and 3) is evident on logical grounds, whether by limiting the need for TW or by making the required TW as generally available as possible. Scores that are lowered because examinees do not comprehend some aspects of complex item formats, or because examinees are reluctant or fearful of using partial information, do not reflect the aptitudes being tested as satisfactorily as scores less influenced by these methods factors. Similarly, the positive effect on test validity of helping students overcome problems in test-taking confidence and efficiency (components C-1 and 2) should be self-evident.

The second of the ultimate objectives for future research is not to discourage concern and efforts regarding test-preparedness, but rather to foster realistic understanding and expectations regarding likely outcomes of STI and ITI by encouraging appropriate emphases and expectations regard-

ing such instruction. Appropriate emphases and expectations may be encouraged by informing students and educators of the appropriateness of STI where needed for score components A-2, 3, and 4 (increasing scholastic aptitude performance), B-1 (filling in important gaps in assumed knowledge or skills), B-2 and 3 (providing general and specific TW), and C-1 and 2 (helping examinees develop test-taking confidence and efficiency). Inappropriate STI or ITI emphases, expectations, and activities may be discouraged primarily by informing students and educators of the limitations of special instruction corresponding to the components of test scores and to related examinee characteristics. This could begin by noting that for most students score component A-1 is by far the largest and is by definition not subject to STI effects, and by noting that because of component D-1 (error variance) any program of STI or ITI may result in a certain percentage of substantial score gains that are attributable entirely to chance and do not, therefore, constitute bona fide evidence of STI effects. Attention can then be directed to those score components that may be influenced by STI but for which such effects are necessarily subject to strong limitations. The effects of STI addressed to component A-2 (review), for example, are subject to two limiting factors. First, "review" presupposes that relevant material had already been learned earlier, and second, the effects for a given examinee are necessarily limited by the extent of his or her need for review. Similar limitations related to "readiness" for STI and the degree of need for it apply to components A-3 and 4. STI effects for score component B-1 are limited by the number of test items calling for the required knowledge or skill, and by the degree to which the examinee is lacking in these skills. Components B-2 and B-3 are limited by the degree of test naiveté to be overcome, as well as the need for TW that the test imposes. On the latter point, for example, an examinee who can answer most test questions correctly with confidence has little need for an effective strategy for guessing on the basis of partial information; the converse is true, of course, for the examinee who is only partially informed on a large percentage of the test questions. STI for components C-1 and C-2 is limited in its effects primarily by the extent to which examinees are being handicapped by a lack of confidence and efficiency in test taking.

Perhaps the clearest instance of STI limitations that can be pointed out is instruction consisting almost entirely of drill on sample test questions. Such instruction is not only academically unsound but misses most of the avenues for having a meaningful effect on test scores. It entirely bypasses components A-2, A-3, A-4, and B1 and deals only peripherally with the TW components B-2 and B-3. Furthermore, it is unlikely to have more than a very modest effect on C-1, since confidence can best be built on a realization of increased competence in coping with the informational and TW requirements of the test, or on C-2.

Before leaving the topic of the limitations of STI, it is useful to address the paradox that despite these limitations, an examinee's problems with one or more of the score components may be such that appropriate STI could result in a very large score gain. The resolution of the paradox is in the realization that the limits are in the form of a "ceiling" effect, but that there is no equivalent "floor" effect. For example, some examinees may be so lacking in test-taking confidence that they "bomb" on the SAT, and remedying this may appropriately result in meaningfully and appropriately large score gains. However, the ceiling effect is such that: (1) the STI cannot yield a test score higher than that warranted by developed aptitude; and (2) such large gains can only

occur for those examinees who were initially severely handicapped by poor test-preparedness. To put it more generally, those students who are already well test-prepared will have little gain from STI, whatever its quality and duration, their own levels of motivation, and so on.

The third and ultimate objective for the recommended research would be derived in the process of realizing the first two. This objective is to gain a more basic understanding of the processes involved in test-taking that contribute to the test scores. Such an understanding can provide a good foundation for a possible evolution in aptitude testing, and could assist in providing information for diagnostic and placement purposes, rather than for admissions decisions only.

Immediate objectives. Recommended immediate objectives of future research regarding STI and TW would be to study systematically the effects of STI (or ITI) directed to the several components of SAT-M and SAT-V test scores, taking into account selected characteristics of examinees, test items, and special instruction. Examinee characteristics of most direct interest would be those related to the several test score components. Examples of these, for which measures before and following STI would be desirable, are as follows. For components A-2 and A-3, the level of mastery of skills such as computing ratios and proportions; for A-4, observations of item-answering processes, and facility in locating required information by scanning reading passages; for B-1, measures of information and skills (such as understanding the test directions) assumed in those taking the SAT; for B-2 and B-3, degree and kinds of TW and test-naiveté, including those involved in guessing behavior; and for C-1 and C-2, indices of levels of confidence and efficiency in test-taking. Among the item characteristics of interest would be item format, difficulty, fineness of distinction between the distractor and the keyed choice, and so on. Characteristics of the STI would include the instructional materials used and the conditions under which they were used. For score components A-2, 3, and 4, the use or nonuse of such materials as mathematics review and vocabulary building textbooks would be of nterest. Similarly, instruction for components B-1, 2, and 3 might be examined for differences associated with the use or nonuse of test familiarization materials more or less resembling the SAT descriptive booklets. For components C-1 and 2, effects resulting from taking a practice test, particularly one under conditions closely paralleling the SAT (such as the PSAT) would be of interest. Whatever the instructional materials, other variables of interest would be whether the STI was undertaken alone, through a tutor, in a more or less typical classroom setting, or in commercial coaching sessions or their equivalent.

DESIGN CONSIDERATIONS

The bulk of the studies reviewed in this survey have contributed little toward providing information that is either broadly generalizable or cumulative. This may be attributed in part to the considerable complexities of the questions involved, in part to the provision of STI that was loosely structured and monitored, with little certainty of exactly what was provided, and in part to a tendency to consider the results in an overall way with scant attention paid to systematic differences among score components, examinees, test items and instruction. It would seem important, therefore, that future research on STI and TW as they apply to the SAT should be given strong design consideration. This does not mean, of course, that pilot studies should be excluded, nor does it mean that only large-scale, costly

studies should be undertaken. What it does mean is that future studies of any magnitude should be tightly designed and fit into a matrix of interrelated studies that will collectively shed considerably more light on these questions and give more information than is now available. These studies, directed to the objectives presented above, can be given strong design characteristics by partitioning between STI effects and practice and growth, by partitioning among score components and examinee, item, and instructional characteristics, and by the systematic gathering and use of detailed pretest and posttest information.

Pretest and posttest data should be collected but not simply for overall test scores averaged over all examinees. For a given study, contrasts in STI effects associated with item format or with other item characteristics may well be desirable. Such contrasts should be designed into the study, with consideration given to adequate sampling of an item pool for each category of items that is of interest, and in particular they should have a large enough number of items in each category for meaningful and stable score differences to be demonstrated where appropriate. Pretest and posttest measures on examinee variables are desirable across a fuller array of score components than have generally been used. This is particularly true for such aspects of TW as guessing strategies and RT tendencies. Ideally, a test-preparedness profile over the several score components could be of great value.

RECOMMENDATIONS FOR SAT-M RESEARCH

The Pike and Evans study of instruction for the SAT-M demonstrated very clearly that meaningful STI effects, both overall and differentially by item format, can be obtained if the instruction is well designed and if it covers a sufficient range of SAT-M score components, including mathematical content (A-2, A-3, A-4, and B-1) and general and specific TW (B-2 and B-3). Recommended future research would be in the form of a coordinated series of studies, replicating that of Pike and Evans but differing primarily in an emphasis on partitioning the effects according to selected examinee, item, and instructional characteristics as they apply to selected test components.

The outcome of a set of such studies would include: (1) an extension of the study to effects observed for inner-city students or others likely to be at lower levels of developed mathematical aptitude; (2) essentially a replication, but dropping data sufficiency items and differentiating between the content score components (A-2, A-3, A-4 and B-1) and the TW score components (B-2 and B-3). The use of diagnostic pretests and posttests of basic content skills and knowledge, and of TW abilities and attitudes, would be an integral part of the study design. Over the set of studies there would also be differentiation based on selected item characteristics within item format, differentiating between instruction provided by self-study and instruction provided under classroom supervision.

There are, of course, many ways of dividing the work into a set of related studies. Among the possible studies that would seem most to warrant current consideration are the following: (1) an extension of some part of the Pike and Evans research to different examinee groups, in particular inner-city students or students identified as highly "math anxious"; (2) a study focused primarily on TW score components--it is in this realm that the questions of fairness and test validity are most problematic; and (3) a study directed primarily to relevant mathematical content. Either of the last two could be profitably expanded or followed up by a companion study

that would allow a comparison between self-study and study provided in the classroom.

RECOMMENDATIONS FOR SAT-V RESEARCH

There is no SAT-V counterpart to the Pike and Evans SAT-M study upon which to build a series of subsequent studies. There are, however, two studies that can serve in conjunction with the research objectives and design considerations already outlined to give some direction to future SAT-V research.

In the Pallone study, instruction was carefully designed and monitored and was directed to the full array of STI or ITI score components. As with the Pike and Evans study of the SAT-M, the resulting score gains were both pragmatically and statistically significant. The fact that only a few subjects in a single school were involved, with no control subjects, severely limits the generalizability of the findings and fails to allow a partitioning between instructional effects and those attributable to growth and practice.

The Whitely and Dawis study was limited to STI for verbal analogies. Within that constraint its strong design makes it a good study upon which to base some of the decisions for subsequent research plans for the SAT-V. The subjects were 184 students randomly drawn from two high schools and randomly assigned to treatment and control groups. The instruction was carefully designed and administered, and in the sense that the underlying skill of educing relationships is one that can be taught and may be considered a content skill, the instruction covered both content and TW score components. Resulting score gains were statistically significant and were also pragmatically relevant to the SAT-V if STI effects for component test scores, as well as for overall scores, are considered. The basic limitation on extrapolating from the Whitely and Dawis findings to the analogies part of the SAT-V is that vocabulary was kept at an unusually low level in their study. The vocabulary requirements of SAT-V analogies of above average difficulty often include words that are rather difficult because they are needed to test the ability to educe more subtle relationships.

The current status of firm information regarding possible STI effects for the SAT-V is particularly problematic. Despite the importance of SAT-V scores, the issues of fairness and validity tied to possible STI effects, and the existence of a marked discrepancy between studies reporting negative findings and those reporting meaningful SAT-V score gains, no study exists that seriously tests the strong assertions noted above that the SAT-V is apparently impervious to the effects of periods of special instruction even as long as a year. As noted above, the general belief that the SAT-V must be considerably less susceptible to STI than the SAT-M does not hold up in a comparison of SAT-M and SAT-V effects in the several studies directed to both. Furthermore, several years have passed since Pike and Evans demonstrated STI effects for the SAT-M in 1972, and the Pallone study in 1961, which reported large STI effects for the SAT-V. What appears to be needed is a study or set of studies using STI generally similar to that provided by Pallone but modified to meet the objectives and design requirements described above. The effects of STI should be studied in a manner that would allow a partitioning of results according to test score components as they relate to specified examinee, item, and instructional characteristics. As was recommended for the SAT-M, a partitioning between content-related score components (A-2, A-3, A-4, and B-1), and TW components (B-1 and B-2) would be desirable. Among examinee characteristics of

interest would be variables related to quality of education, and perhaps a group having the verbal equivalent to "math anxiety" or "math aversion." Most basic among item characteristic partitionings would be that of examining STI effects separately by item format: reading comprehension, analogies, antonyms, and sentence comprehension. In doing so, instruction would be tailored to content-related and TW score components as manifested by each item format. Again, the use of diagnostic pretests and posttests of basic content skills and knowledge, and TW abilities and attitudes, would be an integral part of the study design. Over the set of studies, information would also be gathered comparing self-study to study in classroom settings.

References

STUDIES OF SHORT-T[illegible] INSTRUCTION (STI) AND
INTERMEDIATE-TERM [illegible]RUCTION (ITI)

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976.

Carroll, J. B. Possible directions in which College Board tests of abilities and learning capacities might be developed. In College Entrance Examination Board, Report of the Commission of Tests: II. Briefs. New York: College Board, 1970.

Coffman, W. E., and Neun, M. E. Effects of an accelerated reading course on SAT-V scores. Research Bulletin 66-11. Princeton, N.J.: Educational Testing Service, 1966.

College Board. Effects of coaching on Scholastic Aptitude Test scores. New York: College Board, 1965, 1968.

College Board. Report of the Commission on Tests: I. Righting the balance. New York: College Board, 1970.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Downey, G. W. Is it time we started teaching children how to take tests? The American School Board Journal, January 1977, pp. 26-31.

Dyer, H. S. Does coaching help? College Board Review, 1953, 19, 331-335. (a)

Dyer, H. S. Scholastic Aptitude Test Coaching Study Data: Appendix to the article "Does Coaching Help?" New York: College Board, 1953. (b)

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Evans, F. R., and Pike, L. W. The effects of instruction for three mathematics item formats. Journal of Educational Measurement. Winter 1973, 10(4), 257-272.

Frankel, E. Effects of growth, practice and coaching on Scholastic Aptitude Test scores. High Points, January 1960, 34-45. (a)

Frankel, E. Effects of growth, practice and coaching on SAT scores. Personnel and Guidance Journal, 1960, 38, 713-719. (b)

Fremer, J., and Chandler, M. O. Special studies. Chapter VI in W. H. Angoff (Ed.), The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Board, 1971.

French, J. W. An answer to test coaching: Public school experiment with the SAT. College Board Review, 1955, 27, 5-7. (a)

French, J. W. The coachability of the SAT in public schools. Research Bulletin 55-26. Princeton, N. J.: Educational Testing Service, 1955. (b)

French, J. W., and Dear, R. E. Effect of coaching on an aptitude test. Educational and Psychological Measurement, 1959, 19, 319-330.

Jacobs, P. I. Effects of coaching on the College Board English Composition Test. Educational and Psychological Measurement. 1966, 26, 55-67.

Lass, A. H. Unpublished report. Brooklyn, N. Y.: Abraham Lincoln High School, 1958.

Loret, P. G. A history of the content of the Scholastic Aptitude Test. Test Development Memorandum 60-1. Princeton, N.J.: Educational Testing Service, 1960.

Marron, J. E. Special test preparation, its effects on College Board scores and the relationship of effected scores to subsequent college performance. Unpublished manuscript, 1965.

McCarthy, R. P. Personal communication. Longmeadow, Mass., 1976.

Moore, J. C. Test-wiseness and analogy test performance. Measurement and Evaluation in Guidance, Winter 1971, 3(4), 198-202.

Pallone, N. Effects of short- and long-term developmental reading courses upon S.A.T. verbal scores. Personnel and Guidance Journal, 1961, 39, 654-657.

Pike, L. W. Implicit guessing strategies of GRE-Aptitude examinees grouped by sex and ethnicity. Unpublished manuscript, 1978.

Pike, L. W., and Evans, F. R. Effects of special instruction for three kinds of mathematics aptitude items. Research Report No. 1. New York: College Board, 1972.

Roberts, S. O., and Oppenheim, D. B. The effect of special instruction upon test performance of high school students in Tennessee. (College Board Research and Development Reports, RDR-66-7, No. 1.) Research Bulletin 66-36. Princeton, N. J.: Educational Testing Service, 1966.

Stanley, J. C. Reliability. Chapter 13 in R. L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Vernon, P. E. Symposium on the effects of coaching and practice in intelligence tests. V. Conclusions. British Journal of Educational Psychology, 1954, 24, 57-63.

Whitely, S. E., and Dawis, R. V. Effects of cognitive intervention on latent ability measured from analogy items. Journal of Educational Psychology, 1974, 66(5), 710-717.

Whitla, D. K. Effect of tutoring on Scholastic Aptitude Test scores. Personnel and Guidance Journal, 1962, 41, 32-37.

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

STUDIES OF TESTWISENESS (TW)

Alker, H. A., Carlson, J. A., and Hermann, M. G. Multiple-choice questions and student characteristics. Research Bulletin 67-52. Princeton, N.J.: Educational Testing Service, 1967.

Bath, J. A. Answer-changing behavior on objective examinations. The Journal of Educational Research, 1967, 61, 105-107.

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

Connolly, J. A., and Wantman, M. J. An exploration of oral reasoning processes in responding to objective test items. Journal of Educational Measurement, 1964, 1(1), 59-64.

Crehan, K. D., Koehler, R. A., and Slakter, M. J. Longitudinal studies of test-wiseness. Journal of Educational Measurement, Fall 1974, 11(3), 209-212.

Diamond, J. J., and Evans, W. J. An investigation of the cognitive corre-

lates of test-wiseness. Journal of Educational Measurement, Summer 1972, 9(2), 145-150.

Diamond, J. J., and Evans, W. J. The correction for guessing. Review of Educational Research, 1973, 43, 181-191.

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

Gentile, J. R. Toward an experimental analysis of reasoning on the Scholastic Aptitude Test: A pilot study. Research Memorandum 66-26. Princeton, N.J.: Educational Testing Service, 1966.

Gentile, J. R. Sociocultural level and knowledge of definitions in the solution of analogy items. American Educational Research Journal, 1968, 5, 626-638.

Gentile, J. R., Kessler, D. K., and Gentile, P. K. Process of solving analogy items. Journal of Educational Psychology, 1969, 6, 494-502.

Jacobs, S. S. Answer changing on objective tests: Some implications for test validity. Educational and Psychological Measurement, 1972, 32, 1039-1044.

Lord, F. M. The effect of random guessing on test validity. Educational and Psychological Measurement, 1964, 24, 745-747.

Lord, F. M. Formula scoring and number-right scoring. Journal of Educational Measurement, 1975, 12, 7-11.

Lynch, D. O., and Smith, B. C. Item response changes: Effects on test scores. Measurement and Evaluation in Guidance, 1975, 7(4), 220-224.

Millman, J., Bishop, H., and Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Mueller, D. J., and Schwedel, A Some correlates of net gain resultant from achievement test items. Journal of Educational Measurement, 1975, 12, 251-255.

Mueller, D. J., and Wasser, V. Implications of changing answers on objective test items. Journal of Educational Measurement, 1977, 14, 9-13.

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

Pyrczak, F. Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. Reading Research Quarterly, 1972, 8, 62-71.

Pyrczak, F. Passage-dependence of items designed to measure the ability to identify the main ideas of paragraphs: Implications for validity. Educational and Psychological Measurement, 1974, 34, 343-348.

Slakter, M. J. Risk taking on objective examinations. American Educational Research Journal, 1967, 4, 31-43.

Slakter, M. J. The penalty for not guessing. Journal of Educational Measurement, 1968, 5, 141-144. (a)

Slakter, M. J. The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 1968, 5, 217-222. (b)

Slakter, M. J. Generality of risk taking on objective examinations. Educational and Psychological Measurement, 1969, 29, 115-128.

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

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

Thorndike, R. L. Editor's note: The problem of guessing. In Educational Measurement. Washington, D. C.: American Council on Education, 1971.
Tobias, S. Math anxiety and avoidance. Seminar presented orally at Educational Testing Service, Princeton, N.J., May 1977.
Vernon, P. E. The determinants of reading comprehension. Educational and Psychological Measurement, 1962, 22(2), 269-286.