Review of Educational Research
Fall 1990, Vol. 60, No. 3, pp. 373–417

Coaching for the Scholastic Aptitude Test: Further Synthesis and Appraisal

Betsy Jane Becker
Michigan State University

Analyses of results of studies documented in 23 reports on coaching for the Scholastic Aptitude Test (SAT) indicated that, on average, coaching can help increase SAT scores. However, considerable variability in results for 48 studies reflected the fact that not all coaching is necessarily effective and that all studies of coaching do not provide similar views of coaching's effectiveness. Results of published and unpublished studies were analyzed separately. Characteristics related to the magnitudes of coaching effects included date of publication of the study, whether the study was sponsored by the Educational Testing Service and used as a comparison group, whether instruction included test practice and attention to test-taking skills, and whether homework was assigned to students. Coaching effects were stronger for the SAT Mathematical subtest. Published comparison studies gave consistent results with coached groups exceeding controls by 0.09 standard deviations on SAT-V and 0.16 on SAT-M. Studies of coaching were found to be rather poorly reported and designed without much attention to the issues discussed in reviews of the coaching literature.

During the 1980s the topic of performance on the Scholastic Aptitude Test (SAT) attracted much attention in academic circles (e.g., Benbow & Stanley, 1980; Powell & Steelman, 1984) as well as in the popular press (e.g., Owen, 1985; Staples, 1985). Because SAT scores are widely used in college admissions decisions, potential examinees (often urged on by their parents) are highly motivated to perform as well as they possibly can. Strong public interest is reflected in the proliferation of commercial coaching schools, computerized coaching programs (e.g., Owens, 1983; Staples, 1985), and school policies on SAT preparation (National School Boards Association, 1984).

Scholarly interest in the effectiveness of coaching for the SAT extends back over 30 years (e.g., Dyer, 1953). The Educational Testing Service (ETS), which develops and administers the SAT, has long claimed that coaching does little to raise SAT scores (e.g., College Entrance Examination Board [CEEB], 1983, p. 10); many of the studies of coaching conducted by ETS seem to support this claim. Yet controversy still surrounds the question of SAT coaching.

Since 1980, six reviews of the research on SAT coaching have been published (Bond, 1989; Cole, 1982; DerSimonian & Laird, 1983; Kulik, Bangert-Drowns, &

I would like to thank Christine M. Schram for her assistance with data evaluation and analysis and Larry V. Hedges and the anonymous referees for their suggestions of revisions to earlier versions of this manuscript.

Kulik, 1984; Messick & Jungeblut, 1981; Slack & Porter, 1980). Each review has examined different aspects of the question of coaching effectiveness for slightly different (yet greatly overlapping) sets of studies. This review considers these studies and more recent studies using an alternative measure of effect (the standardized mean-change measure) for synthesizing pretest-posttest research (Becker, 1988).

Use of the standardized mean-change measure together with a generalized least-squares (GLS) approach for modeling multivariate study outcomes[1] makes possible analysis of the following questions not directly addressed in previous reviews:

1. What are the relative contributions of characteristics of the subjects, the coaching interventions, and the studies themselves to the magnitudes of coaching effects, and which of these characteristics are confounded with one another?
2. Is the effect of coaching the same for the Mathematical and Verbal sections of the SAT?
3. Can a "model" or set of predictors be found that explains the variation in the results of these coaching studies?
4. Have any studies produced coaching effects that appear unusual or extreme in light of the outcomes predicted by models of coaching-study results (mentioned in Item 3)?

This synthesis more comprehensively examines questions addressed in past reviews, allowing for comparison of the simultaneous relative contributions of predictors of the results of the coaching studies. Also, more studies are included. Additionally, a new approach is used to include the results of studies without control groups: Nonexistent control-group results are represented by (imputed as) the average results for the existing control groups. Finally, the GLS analysis accounts for dependencies between outcomes, resulting from the interrelatedness of SAT-Mathematical and SAT-Verbal scores and from the use of shared control groups. This kind of comprehensive synthesis capitalizes on the strengths of quantitative-reviewing procedures and reveals unanswered questions about coaching that result from the quasi-experimental (survey-like) nature of research-synthesis data.

The remainder of this review is structured as follows. First, the issues addressed in past reviews of the effects of coaching on the SAT are discussed. In this context some characteristics of studies of SAT coaching are discussed, but individual studies are not critiqued in detail in order to avoid redundancy with the work of Pike (1978), Slack and Porter (1980), and Messick and Jungeblut (1981). The second section is an outline of the methodology used in this synthesis; the standardized mean-change measure for SAT coaching, the coding of study features, and the statistical analysis are each discussed. Results of the analyses are found in the third section. In the final section the results are related to those of past reviews, and implications of the results are drawn about the efficacy of SAT coaching and about future studies of coaching and the SAT.

Issues in SAT Coaching

This section contains an overview of issues identified in reviews of studies on coaching for the SAT and other ability tests. The issues are drawn primarily from five reviews of SAT coaching published during the 1980s. These are not the only relevant reviews (e.g., reviews of other topics sometimes discuss SAT coaching), but they are the most recent and they identify the ongoing issues in the coaching literature.

Definitions of Coaching and the Content of Coaching Interventions

As Cole (1982) has pointed out, the word *coaching* "is used to refer to a wide variety of test preparation activities undertaken by individuals in an attempt to improve test scores" (p. 389). Her brief definition views coaching as "instructions given in preparation for taking a test that are designed to elicit maximum performance" by the coached examinee (1982, p. 391).

Pike (1978) presented a conceptual framework for understanding coaching in terms of components of students' test scores. A student's score is considered to be composed of "true-score" components (including developed knowledge, relevant analytical skills, extent of overlearning, etc.), primary and secondary test-specific components, and random error. Primary test-specific aspects include general test wiseness and match between examinee ability and test content, whereas secondary test-specific components include such factors as confidence and efficiency. Much coaching is aimed at affecting students' scores by modifying their test-specific knowledge.

Bond (1989) simplified Pike's elaborate 10-component conceptualization into a model with 3 components—a true-score (or alpha) component, a test-specific (beta) component, and random error. Both alpha and beta abilities may be the target of coaching interventions, although Bond noted that most standardized tests are constructed so that beta abilities constitute a minimal part of examinees' scores.

The extent to which coaching attempts to modify the alpha component of test performance relates closely to the amount of content-relevant instruction presented. Kulik et al. (1984) examined differences between three sets of SAT coaching programs: short-term programs of test-taking orientation and practice, longer term programs of drill or "cramming" on test items (which were slightly more effective), and programs of "instruction in broad cognitive skills" (1984, p. 181). ETS has long claimed that "longer-term preparation that develops skills and abilities can have greater effect" (CEEB, 1983, p. 10) on SAT scores than simple drill and practice on items.

Kulik et al. (1984) also coded numerous specific characteristics of the coaching instruction, including whether coaching was conducted by schools or commercial coachers, whether students were instructed in test-wiseness skills, anxiety-reduction techniques, and specific content skills, whether students practiced particular item formats, and whether the coaching program was new or had been field tested. None of these predictors was significantly related to SAT coaching effects, although most differences were predictable (e.g., programs with test-wiseness, test-anxiety, or content-instruction components showed larger effects than programs without those components).

Most other reviews have not systematically or empirically examined specific aspects of the content of coaching interventions. Both Slack and Porter (1980) and Messick and Jungeblut (1981) dealt with these issues only discursively. DerSimonian and Laird (1983) based their sample of studies on those reviewed by Messick and Jungeblut (1981) and Slack and Porter (1980). They did not provide a definition or description of coaching.

The Role of Program Duration

Slack and Porter's (1980) controversial review initiated much of the continuing interest in SAT coaching. On the basis of data from 10 studies published prior to

1968, they claimed that "there is ample evidence that students can successfully train for the SAT and that the more time students devote to training, the higher their scores will be" (1980, p. 164). This conclusion implies that Slack and Porter specifically investigated differences between short- and long-term programs. However, as Jackson (1980) noted, they did not. The conclusion was drawn from a narrative discussion of all results.

Messick and Jungeblut (1981) empirically studied the relationship between number of student contact hours during coaching and size of coaching effects across studies using regression analyses.[2] Logarithmically transformed contact hours appeared linearly related to score effects, and relationships were positive in all cases. Although they noted that contact hours were confounded with methodological features of the studies (such that uncontrolled studies tended to be longer), Messick and Jungeblut reported that for both the mathematical and verbal subtests a "threshold" amount of 3 hours of coaching was needed to attain any positive gain in SAT scores beyond control-group gains.

Messick and Jungeblut reported a steeper slope coefficient for the regression of SAT-M gains (vs. SAT-V gains) on log-time, suggesting a greater relative impact of coaching on mathematics scores. They concluded that improvement of SAT scores "is a function of the time and effort expended and that each additional score increase may require increasing amounts of time and effort, probably geometrically increasing amounts" (1981, p. 215). Short-term instruction, then, would be expected to produce only small-scale gains.

Kulik et al. (1984) categorized 14 studies of coaching effectiveness into average-length (3 to 9 hours) or long-duration (more than 9 hours) programs. Although coached-group advantages were 0.08 standard deviations for average programs and 0.16 standard deviations for long programs, they did not differ significantly. This differs slightly from Messick and Jungeblut's findings. However, only controlled studies were reviewed by Kulik and his coworkers, whereas Messick and Jungeblut's review included several longer, uncontrolled studies. The lack of correspondence in findings may have resulted because the syntheses considered different sets of studies.

Study Quality

Some of the controversies in the literature on coaching relate less to the nature of the coaching intervention and more to the coaching research itself. One such issue concerns the quality of empirical studies of coaching effects.

Slack and Porter (1980) discussed the *results* from 10 reports of coaching studies in detail, but they paid little attention to the question of study design. Jackson's (1980) critique of their review focused on this point.

Messick and Jungeblut (1981) systematically critiqued the methodologies used in studying coaching effects, focusing on the lack of comparison groups, the use of nonequivalent controls, and related problems of differential selection and maintenance of subject motivation. Also, they pointed out the great number of studies that used special administrations (or forms) of the SAT, which could introduce other biases. Individual design flaws were compounded for studies that suffered several problems at once. Additionally, Messick and Jungeblut noted the imprecision in many study results due to small numbers of coached students.

Although they organized their discussion of studies on the basis of design characteristics, Messick and Jungeblut did not present overall results separately for studies with different designs. They stated that "for numerous reasons, including the diversity of design limitations and the differences in sample sizes, it is difficult to compare results across these studies in a meaningful way" (1981, p. 202).

Bond's (1989) treatment of the study-design issue closely followed that of Messick and Jungeblut. Bond's collection of studies differed slightly from Messick and Jungeblut's, and he did summarize results for comparison and uncontrolled studies separately. He reported very large adjusted gains for uncontrolled studies of 38 points for Verbal and 54 points for Mathematical subscores. Smaller average advantages for coaching (of 9.1 SAT-V and 13 SAT-M points) were reported for comparison studies.

DerSimonian and Laird (1983) incorporated information about study design into their quantitative synthesis. They examined all of the studies reviewed by Slack and Porter and Messick and Jungeblut, using a random-effects conceptualization of the coaching outcomes. That is, they considered observed variability in the outcomes of coaching studies to have two components, one due to "true" differences in the effects of coaching and another due to sampling error.

DerSimonian and Laird evaluated three models for "true" differences in coaching effects: the model of a single common coaching effect, the model of investigator-related differences (due to intercorrelated results for multiple studies done by single investigators), and the model of design-related differences in coaching effects. Table 1 shows a summary of the estimated gains in SAT scores for the three models.

The most striking of DerSimonian and Laird's findings may be the wide disparity in mean gains shown for the three study designs. For both math and verbal scores, uncontrolled studies showed gains three to five times larger than comparison studies. This, coupled with the inherent difficulty in evaluating the results of uncontrolled and nonequivalent-comparison-group studies, prompted DerSimonian and Laird to state that "it appears that the benefits of coaching are indeed

TABLE 1

Mean gains on the SAT (and standard errors) for three models estimated by DerSimonian and Laird (1983)

Model	Mean point gains	
	Verbal	Math
Common effect	22.8	21.1
	(5.3)	(4.4)
Common effect adjusted for intercorrelation	19.3	17.7
	(6.6)	(5.1)
Study design effects		
Uncontrolled studies	40.6	53.8
	(10.1)	(5.1)
Controlled studies	15.3	15.6
	(5.5)	(2.6)
Matched or randomized studies	10.1	9.8
	(3.5)	(3.8)

negligible, only about one-tenth of the population standard deviation" (1983, p. 13).

DerSimonian and Laird also discussed the consequences of the confounding of type of study design and type of coaching program. Type of study design (uncontrolled, controlled, and matched or randomized) was associated with both type and selectivity of school and with number of student contact hours. Thus, their results and Messick and Jungeblut's (1981) findings (that coaching effects relate positively to number of student contact hours) are alternative but competing explanations for the patterns of coaching-study outcomes. DerSimonian and Laird therefore concluded that evidence was not sufficient to attribute large positive SAT gains to the effects of coaching *programs*.

Kulik et al. (1984) dealt differently with the issue of study quality. They reviewed only studies with control groups[3] and investigated the importance of the use of randomization. Contrary to DerSimonian and Laird's findings, Kulik et al. found larger coaching effects in randomized studies. Coached students showed average advantages of 0.21 standard-deviation units in four randomized studies, versus advantages of 0.12 standard deviations for coached students from 10 nonrandomized studies.

The interpretation of results obtained using various research designs is discussed in many texts and handbooks on research methods (e.g., Campbell & Stanley, 1963). Choosing between these designs typically involves making compromises between aspects of internal and external validity. For example, studies of students who attended (on their own) commercial coaching schools have high external validity if the researcher wants to infer to the population of students who seek coaching. But inferences about the efficacy of the coaching intervention in the general population (e.g., about the consequences of making coaching mandatory) cannot be based on such studies because of the problem of self-selection of treatment groups. Motivation, ability, and so forth, may also differ for students seeking coaching, as discussed below. Conversely, randomized controlled studies have high internal validity, sometimes at the price of low generalizability.

The authors of primary research on coaching have made a variety of different compromises. Reviewers of that research agree that these choices of research design in coaching studies have had consequences for the inferences that can be justified about coaching itself.

Motivation and Selectivity of the Samples

The degree to which students seek coaching (or volunteer to receive it) represents both the students' desires to be coached and (presumably) their wishes to perform well on the SAT. The degree to which control subjects are motivated is not typically investigated. Studies of commercial coaching (e.g., the controversial Federal Trade Commission [FTC] study [1978, 1979]) are particularly problematic because they focus on highly motivated subjects who are willing (and able) to pay for costly coaching. Such coached subjects, and their comparison-group counterparts, may have also sought help in test preparation in addition to that provided in the studies.

The extent to which the subjects of a coaching study are selected from special schools or unusual groups of students may also have an impact on the effects of coaching. Selectivity of samples in the coaching studies varied widely.[4] Selectivity is a gross measure of student ability and past educational quality, and also reflects

the degree to which students might be expected to improve upon retesting with or without coaching, as discussed below. It may also be a proxy for differences in variability (with more selective samples being less variable) or for differences in coaching treatments (due, for instance, to higher funding levels in more selective private schools).

Most reviewers have not examined these two aspects of coaching samples systematically. Often, reviewers have treated the nature of samples discursively, but have not included sample selectivity and motivation in their analyses.

Kulik et al. (1984) did code ability level of the samples, and found the largest gains (0.20 standard deviations) for low-ability samples. High-ability students showed gains of only 0.06 standard deviations. The authors also compared studies of commercial versus school-based coaching, and found a slight advantage for school-based programs (0.16 standard deviations versus 0.13 for commercial coaching). This latter result is counter to expectations about the effect of student motivation in commercial coaching but is based on only four commercially coached samples, and is not a statistically significant difference.

Gain Due to Retesting

A recurring issue in discussions of uncontrolled studies of coaching concerns the gains that students can be expected to make due to simply being retested. The issue is complicated, because expected gains vary according to the initial score levels of examinees, the duration between first and second testing, and according to which subtest of the SAT is of interest.

Slack and Porter (1980), Messick and Jungeblut (1981), and Bond (1989) all discussed possible adjustments for retesting gains. Slack and Porter suggested using the retesting gains of national samples of students scoring at the same initial levels as coached students. However, Messick and Jungeblut criticized that method because of differences (e.g., in motivation levels and interest) between coached groups (especially when they include volunteers) and national samples.

Bond (1989) discussed research on the effects of growth and retesting on the SAT but noted that further study, attending to both sampling and regression considerations, is still needed before the issue can be well understood. He concluded that SAT-V gains (over 6 months) depend on initial score levels, with students scoring high on initial testing experiencing greater growth. Six-month SAT-M gains, however, were expected to be approximately 15 to 20 points for all examinees.

Score change due to retesting and growth is an even more complicated issue in the coaching literature for several reasons. Students are often tested with nonstandard forms of the SAT (or with the Preliminary SAT [PSAT]), which are given at special administrations of the test. Also, in most cases the exact time interval between the two testing sessions is not reported. Even when comparison groups have been used in coaching studies, they have often been nonequivalent control groups (e.g., from different schools than the coached students). Thus, the gains expected from growth and retesting may not be equivalent even for groups *from the same study.*

Differential Effects of Coaching on SAT-M and SAT-V

A similar issue concerns whether coaching itself is differentially effective at changing scores on the two primary SAT subtests. Most reviews have separated

results for the Mathematical and Verbal sections of the SAT, and several studies of coaching have examined effects on only one SAT subtest (e.g., Alderman & Powers, 1980; Evans & Pike, 1973; Pallone, 1961).

Traditional expectations suggest that mathematics performance should be the easier of the two areas to influence. This idea was partially supported by the analysis of Slack and Porter (1980). Their weighted average effects showed gains of 33 points on SAT-M versus 29 on SAT-V.

Messick and Jungeblut (1981) found that 12 hours was required to achieve a 10-point gain on SAT-V, 4 hours more than was needed to achieve an equal-sized gain on SAT-M. Their extrapolations suggested that students could expect to gain 40 SAT-M points with 107 hours of coaching, but a 40-point gain on SAT-V would require 1,185 hours of coaching (over 7 months of 40-hour work weeks or one full year of schooling with 6.5 hours of classes per day).

Bond (1989) reported stronger coaching effects for SAT-M in both uncontrolled and comparison studies. He also noted that coaching on certain aspects of the above-mentioned alpha components (especially those related to review of subject matter) might be more likely to produce gains in the area of mathematics.

DerSimonian and Laird's review is the only one that fails to substantiate stronger coaching effects for SAT-M across all studies considered. As Table 1 shows, SAT-M gains were 13 points larger than Verbal gains for uncontrolled studies but were close to or slightly less than SAT-V gains for comparison studies.

The review by Kulik and his coauthors (1984) is the only recent review that did not look for differential coaching effects. The authors reported only one effect size for each study, even when both SAT-M and SAT-V had been tested. In such cases the single effect represented either the average of SAT-M and SAT-V effect sizes or an effect calculated for total SAT scores. Thus, their review did not address the issue of differential effects on SAT-M and SAT-V scores.

None of the reviews provided statistical tests of differences between coaching effects for Math and Verbal outcomes.

Summary

Although there is considerable variability in the results of coaching studies, reviews of those studies show a remarkable amount of consistency. Several common conclusions can be stated:

1. The results of studies of coaching effectiveness vary widely.
2. Much of the variation in results arises from studies without comparison groups.
3. Across all studies, the magnitudes of coaching effects relate to study design, as well as to duration of the coaching interventions.
4. Features of study design and of the coaching interventions are confounded in the set of coaching studies.

Method

This section begins with a definition of the standardized mean-change measure (Becker, 1988) and a description of the computation of the index of coaching effectiveness. Study retrieval and the coding of study characteristics are detailed, as is the analysis of data.

Standardized Mean-Change Measure

The measure of effectiveness of coaching used in this review is based on the standardized mean-change measure outlined by Becker (1988). The standardized mean change is computed for each subgroup in a study (e.g., for each coached and each uncoached group). A study with one coached and one uncoached group would have two standardized mean changes for each outcome, each computed as the change for one group in mean performance from pretest to posttest, divided by the posttest standard deviation. Denoting the standardized mean changes as g^C for the coached group and g^U for the uncoached group, then

$$g^C = \frac{(\bar{Y}^C - \bar{X}^C)}{S_Y^C} \text{ and } g^U = \frac{(\bar{Y}^U - \bar{X}^U)}{S_Y^U},$$

where $\bar{X}^C$ and $\bar{Y}^C$ represent pretest and posttest SAT means for the coached group and S_Y^C is the coached group's posttest standard deviation, and $\bar{X}^U$, $\bar{Y}^U$, and S_Y^U are the analogous statistics for the uncoached sample. Separate standardized mean changes were computed for SAT-M and SAT-V. Total scores were not used in any cases.

The statistics g^C and g^U are slightly biased estimates of the population standardized mean-change parameters, but unbiased estimates d^C and d^U are available (see Becker, 1988, and Appendix C of this paper). The standardized mean change provides a scale-free measure of the amount gained by the average subject in each sample relative to others in the sample. The values d^C and d^U can be interpreted in standard deviation units, in the same way that Glass's (1976) effect size is typically interpreted.

The statistic of interest for examining coaching effects is the difference between the (unbiased) standardized mean changes,

$$\hat{\Delta} = d^C - d^U.$$

Studies that examine both SAT-M and SAT-V performance will have two differences: $\hat{\Delta}^M$ for the SAT-M mean-change difference and $\hat{\Delta}^V$ for SAT-V.

The statistic $\hat{\Delta}$ has several advantages over Glass's effect size for examining the coaching literature. First, some studies of coaching present results for several coached groups but only one control group. Thus, indexes of coaching effectiveness that compare each of the coached groups to a single control are dependent.[5] The covariances between $\hat{\Delta}$ values that involve a single (common) control group are straightforward, whereas those between effect sizes computed for the same groups are much more complex (because of the shared standard deviations), and consequently have not been described in the statistical literature. Similarly, covariances between math and verbal study outcomes between studies with shared control groups are also simple and easily derived.

Third, the fact that the standardized mean change is a simple contrast between an index for a coached and a control group suggests a new approach for including studies without control groups. Methods similar to those used for handling *missing* data (e.g., Rubin, 1987) can be applied to impute control "results" for the studies without real control groups. Here the average standardized mean changes on SAT-M and SAT-V for all existing control groups were used as proxies for the nonexistent

results. Thus, mean-change differences for studies without control groups were denoted as $\tilde{\Delta}$, and computed as:

$$\tilde{\Delta} = d^{C} - d^{U}$$

where d^{U} represents the weighted average standardized mean change for all control groups on SAT-M or SAT-V. This method relies on the assumption that if control groups had been obtained in the no-control studies, these results would have been similar to those of the existent control samples. The assumption is, of course, impossible to verify. However, some information can be obtained by considering the consistency of results for the existing control groups. This is discussed in the Results section below.

Coding of Study Characteristics

Study characteristics must be considered in any investigation of results of coaching-effectiveness studies in order to evaluate their potential relationships to study results. Study characteristics that differ across studies, but would *not* be expected to relate to study outcomes, are important to consider because these features need to be eliminated as potential explanations of patterns of study results. Because of the survey-like (unstructured) nature of meta-analysis data, several study features may be confounded for a particular set of studies. Messick and Jungeblut (1981) have already noted that several features of coaching effectiveness studies are confounded. Thus, several competing "models" or explanations of patterns of study outcomes may exist. Inclusion of as many study features as possible in analyses of the study results allows those different models to be evaluated.

Nineteen characteristics of the sample, the coaching intervention, and the design and reporting of coaching studies were coded, and are listed in Table 2. Two raters coded the characteristics for each study, and discrepancies between assigned codes were resolved. The percentages of codes on which both raters agreed (before resolution of differences) are a measure of interrater reliability. Percentages ranged from 56% for sample voluntariness to 100% for four variables coded. Only two variables other than voluntariness had reliabilities below 80%—duration of coaching with 79% agreement and presence of test practice with 71% agreement. Problems encountered in coding these variables are discussed below.

Study characteristics. Six variables described the design and reporting of the study itself. Year of publication allowed examination of the question of trends over time, and affiliation of authors with ETS (or reported ETS sponsorship of the study) was also noted. Type of publication (journal articles versus unpublished documents) was also coded.

Sample characteristics. The selectivity of the school or coaching program with which students were associated and the degree to which subjects had volunteered to participate in the coaching program were coded. Table 2 lists the three (ordered) categories for each variable (values are shown in brackets).

Voluntariness was difficult to code because of coaching offered by high schools or college preparatory schools as part of their regular curricula. One major controversy centered on the voluntariness of the Marron (1965) samples. The subjects were exposed to coaching as a part of the *required* curricula of several college preparatory schools, suggesting a coded value of zero. However, the schools

TABLE 2
Characteristics of studies for coaching for the SAT

Characteristic	Values
	Study characteristics
Year of publication	Last two digits of date coded (e.g., 1976 coded as 76)
Type of publication[a]	Articles published in academic journals
ETS authorship[a]	At least one author was affiliated with ETS or the study was reported to be sponsored by ETS
Control-group use[a]	Results of a comparison group were reported (whether randomized or nonrandomized)
Use of matching[a]	Control and coached subjects were matched
Use of randomization[a]	Subjects were randomly assigned to control and coaching groups
	Sample characteristics
Selectivity	Low achievers or underprepared students [0] Public-school students or mix of students from public and private schools [1] College-preparatory-school students or students from selective public high schools [2]
Voluntariness	Participation in coaching was compulsory [0] Participation was easily possible and available at little or no cost (e.g., as a school elective) [1] Participation was totally voluntary, coaching was extra curricular or obtained from a commercial coacher [2]
	Coaching-intervention characteristics
Duration	Length of coaching program in hours
Presence of verbal coaching[a]	Instruction given was relevant to taking of SAT-V (e.g., vocabulary drill, practice with verbal analogies)
Presence of mathematics coaching[a]	Instruction given was relevant to taking of SAT-M (e.g., practice in quantitative item types, drill on mathematics skills)
Presence of item practice[c]	Participants were given practice on sample test items [1] Participants were not given practice on sample test items [2] No information [3]

(continued on p. 384)

TABLE 2 (*continued*)

Characteristic	Values
Presence of test practice[c]	Participants practiced taking complete sample tests [1] Participants did not practice taking complete sample tests [2] No information [3]
Instruction in test-taking skills[c]	Participants were instructed in such skills as budgeting time, guessing strategies, etc. [1] Participants were not instructed on such strategies [2] No information [3]
Extra activities[c]	Participants experienced other activities such as career counseling, selection of colleges, etc. during coaching time [1] Participants experienced no other activities [2] No information [3]
Homework[a]	Participants completed coaching related exercises outside of the coaching course [1] No outside exercises were given [2] No information [3]
Use of computerized instruction[a]	Computerized coaching program or computerized item practice was used [1] No computerized instruction was used [2] No information [3]
Instruction for alpha abilities	Participants were instructed in more than one knowledge area, such as vocabulary, reading comprehension, algebra [2] Participants were instructed in one area [1] Participants received no instruction in these areas [0]
Type of control[b] group	Wait-list or delayed-treatment control group [1] Alternative-treatment control group [2] No information or no control group [3]

Note. Numbers in brackets produced the coded values shown in Table A-2.

[a] These variables were assigned the value 1 if the specified aspect of the coaching intervention or study design was present and 0 otherwise.

[b] This variable was recoded as two dummy variables for the regression analysis.

[c] These variables were recoded 1 if the specified aspect of the coaching intervention was present and 0 otherwise for the regression analysis.

themselves were 1-year post–high school college preparatory schools, and the extent to which attendance at the schools was voluntary is not clear. (These samples were finally given a voluntariness code of 2 based on the available information.)

Coaching-intervention characteristics. Eleven features of the coaching programs were noted: their duration, their emphasis on mathematical and verbal preparation, seven types of coaching instruction or activity, and type of control treatment.

Duration was estimated for half of the sources because of partial or vague information (e.g., the program was reported to have lasted for 20 sessions, but the duration of each session was not given). Dichotomous (0, 1) variables measured each program's emphasis on mathematics and verbal preparation, and the presence of seven specific kinds of coaching activities. The activities included item practice, test practice, test-taking skill instruction, homework activities, instruction or lessons aimed at changing Bond's (1989) alpha component of the student's test score, and other activities (such as anxiety-reduction exercises) often aimed at changing test-specific (beta) knowledge and behaviors. Test practice had a relatively low reliability (71%), but half of the disagreements resulted from the coding of Marron's (1965) study.[6] Computerized instruction does not specifically fit into Bond's framework, and was used in four studies (Coffin, 1987; Curran, 1988; Davis, 1985; Laschewer, 1986).

Several aspects of study design were coded. Use of any kind of comparison group, use of matched coached and comparison subjects, and use of randomized coached and control groups were noted. Finally, the type of control treatment used (i.e., the activities of the control group) was coded when information was available.

Analysis of Outcomes

Differences in the standardized mean changes between coached and control groups (or between results of coached groups and the imputed mean control values) were analyzed using a generalized least-squares regression approach. The generalized least-squares regression analysis was required because the variances of differences in standardized mean changes were not equal across studies and because correlations (or covariances) among certain outcomes were nonzero. The analysis follows that suggested by Raudenbush, Becker, and Kalaian (1987) for Glass's effect size. Examination of residual variation remaining after inclusion of the predictors allowed the assessment of each model's adequacy.

Variance-covariance matrix for study outcomes. General formulas for the variances and covariances of the $\hat{\Delta}$ and $\tilde{\Delta}$ values have been presented by Becker (1988). However, because of the very specific structure of the studies of SAT coaching, formulas for the variances and covariances of study results are presented in Appendix C of this paper. The two sources of intercorrelation (multiple outcomes and shared control results) produced 10 different covariances among the study results. Table 3 is a list of the 10 kinds of interrelationships, and formulas for the 10 covariances are given in Appendix C.

Results

This section presents the results of a series of analyses of the measures of change in SAT performance. The first subsection contains a description of the studies, the

TABLE 3
Ten types of intercorrelations among study results

Label	Descriptions of correlated outcomes
	Between-study correlations
A	SAT-M results of two uncontrolled studies SAT-M results of one uncontrolled and one comparison study
B	SAT-V results of two uncontrolled studies SAT-V results of one uncontrolled and one comparison study
C	SAT-M result of one uncontrolled study with SAT-V result of another uncontrolled study
D	SAT-M result of an uncontrolled study with SAT-V result of a comparison study
E	SAT-V result of an uncontrolled study with SAT-M result of a comparison study
F	SAT-M results for two experimental groups that shared a common control group
G	SAT-V results for two experimental groups that shared a common control group
H	SAT-M results from one experimental group with SAT-V results from a second experimental group, when two groups shared a common control group
	Within-study correlations
I	SAT-M and SAT-V results from an uncontrolled study
J	SAT-M and SAT-V results from a comparison study SAT-M and SAT-V results for an experimental group which used a group shared by at least one other control experimental sample

outcomes of coached and control students considered separately, and an evaluation of the proposed plan to "impute" average control results in place of nonexistent control outcomes. The remainder of the section describes analyses of the data that lead to several alternative "models" for coaching-study results.

Description of the Studies

Study retrieval. Sources were identified via three search procedures. The majority of the sources had been examined in past reviews. This collection was updated through a search of the Educational Resources Information Center (ERIC) database (for 1983–1989). Also, a search was made of the *Comprehensive Dissertation Index* for dissertations listed since 1985.

Many sources provided results on studies of more than one school or coaching program. The schools and samples reported separately in Alderman and Powers (1980), Evans and Pike (1973), and Marron (1965)[7] were kept distinct for the analyses. Results from uncontrolled studies were included, as were results from Lass (1961), which were omitted by Messick and Jungeblut (1981).

Studies by Coffin (1987), Johnson (1984), Kintisch (1979), and Reynolds and Oberman (1987) were included. Dissertations by Burke (1986), Curran (1988), Davis (1985), Keefauver (1976), Laschewer (1986), and Zuman (1988) also pro-

vided data. Each of these has been included in no more than one past review of SAT coaching. A dissertation by Winokur (1983) was not available.

Several sources or studies within sources were not included in the analyses because they either did not examine coaching or did not provide the data needed to compute standardized mean changes. In particular, samples from both Marron (1965) and Roberts and Oppenheim (1966) that did not receive coaching were not included. Data for only School A of the FTC's study of commercial coaching could be included. (Rock's [1980] reanalysis presented descriptive statistics for pretest and posttest SATs only for School A.) Three samples from Johnson (1984) had provided data, but two of the samples had changed between pretest and posttest because of attrition. The data from those samples could not be used to compute measures of change. The short-term study by Coffman and Parry (1967) was omitted because of concern about the half-length SAT-V form used (cf. Messick, 1980, and Bond, 1989).

The 23 reports in this review included results for more than 50 studies (coached versus control group comparisons) and over 75 independent samples (i.e., separate coached and control groups) of subjects. Results of 77 samples from 48 studies were analyzed. Forty-eight samples (about 62%) received some kind of instruction or coaching for the SAT, and 29 samples served as control groups. The total number of subjects in the analysis was 6,870, with 3,710 (54%) receiving coaching and 3,160 not being coached.

The SAT outcome. Even though all of the studies retrieved in the search claimed to be relevant to SAT coaching, a number of studies had used measures other than the SAT as outcomes. In part, this occurs because of practicalities. Most coaching programs last only a few weeks or months, but the SAT is only given on a few specific dates each year. Validity concerns (e.g., mortality problems and issues of maturation) suggest that pretests and posttests should be given just before and after a treatment has been implemented. Consequently, coaches and researchers do not usually use official administrations of the SAT to measure coaching effects.

When the SAT is administered in unofficial settings, nonstandard forms are often used. Some studies used shortened or other modified forms of the SAT (e.g., Evans & Pike, 1973), old (released) or sample forms (e.g., Zuman, 1988), or simulated SATs constructed by faculty and graduate students in measurement (Johnson, 1984). The assumption that these tests have been adequately equated with current forms of the SAT is dubious.

A number of studies also used the Preliminary Scholastic Aptitude Test (PSAT) as an outcome (e.g., Roberts & Oppenheim, 1966), or used the PSAT as a pretest and the SAT as the posttest (Zuman, 1988). Because the PSAT is meant to be comparable to the SAT (e.g., Anastasi, 1988), this practice is not assumed to be problematic. However, because the PSAT is shorter than the SAT, greater scaled-score point gains may be possible with equivalent gains in numbers of items answered correctly.

The study with the most unusual outcome measure is Johnson (1984). The outcomes were SAT forms developed from existing SAT forms by the author and graduate students in measurement. According to Bond (1989), the forms were equated with the SAT score scale using standard equating procedures.

Analyses based on the standardized mean-change measure do not require outcomes to be on the same scale, thus diminishing somewhat the problems associated

with the potentially differing scales of nonstandard SATs. Other questions of similarity (e.g., in difficulty levels or content), however, still remain. If all nonstandard measures used in this literature were strongly equated with the SAT, the results of analyses on points gained would be essentially the same as those based on measures of change.

Eliminating studies that had used any test other than officially administered SATs would have reduced the number of available sources to only 13 (from 23). Thus, these studies were retained in the analysis and were examined individually in all analyses of model misfit.

Characteristics of the studies. Table 4 lists and describes the 48 studies and their subjects, and Table 5 summarizes study characteristics.

Most studies (88%) examined students of average or above-average ability (i.e., samples with selectivity ratings of 1 or 2). Selectivity of samples also appeared related to variability in subjects' SAT scores. More selective samples were, on average, somewhat less variable than unselected samples.[8] Subjects in most studies had some control over their participation in coaching programs. Only three studies examined compulsory coaching, whereas more than half of the studies involved students who had volunteered for extracurricular coaching or had been paid to be coached.

The bulk of studies provided instruction specifically oriented to SAT-Math or SAT-Verbal performance, with 87% of the studies offering SAT-V related instruction and slightly more than half offering SAT-M instruction. Twenty-three of 30 programs (some programs were attended by several samples) provided instruction in both areas. One intervention that claimed to offer no preparation specific to either SAT subtest (i.e., Coffman and Parry's [1967] accelerated reading course) was assigned values of zero for both emphasis indicators.

The 23 reports were published between 1953 and 1988, with an average publication date of 1973 (and a median date of 1976). The "typical" study was done approximately 15 years ago, raising the question of how the SAT, prevailing levels of education (which may relate to aipha knowledge), and general awareness of testing and test-taking skills have changed since the bulk of the studies were conducted. However, 7 reports (30%) have been issued within the last 5 years; thus, the literature is not totally out of date. Less than half (10 of 23) of the sources had studies conducted or sponsored at least in part by ETS.

The average duration of coaching across 46 studies was 35 hours, with a range of from 4.5 to 100 or more hours of instruction. The duration of Pallone's long-intervention was estimated at 100 hours (50 minutes × 5 days/week × 4 weeks × 6 months), and samples from Marron (1965) were also coded as receiving 100 hours of coaching. (This differs from the value of 300 hours assigned to both of these studies by Messick and Jungeblut, 1981). Duration was estimated similarly for studies from Coffin (1987), Dear (1958), Dyer (1953), Frankel (1960), and French (1955). The median duration across the programs that reported duration (15 hours) was imputed for missing duration values in Kintisch (1979) and Lass (1961).

Standardized mean changes. Because not all of the studies had reported on both SAT-M and SAT-V performance, the 48 studies produced 114 nonredundant standardized mean changes and 70 mean-change differences. The majority of these differences represented performance on the SAT-Verbal sections, with 44

TABLE 4
Description of studies

Study	Duration of coaching (hours)	Content of coaching	Type of sample
		Uncontrolled studies	
Coffin B (1987)	25	Computer-assisted instruction	Urban public schoolers
Coffman & Parry (1967)[a]	48	Accelerated reading	Public university freshmen
Johnson (1984)[b]	30	Math and verbal exercises	Low income black students
Marron (1965)			
Samples 1–7	100	College prep instruction	Private prep school boys
Pallone (1961)			
Short-term program	45	Developmental reading	Private prep schoolers
Long-term program	100	Developmental reading	Private prep schoolers
		Nonequivalent comparison studies	
Curran (1988)[c]	6	Computer assisted instruction	Parochial schoolers
Dear (1958)[a]	10	Math and verbal exercises, homework, and test practice	Students from public and private, coed and same-sex schools
Dyer (1952)[a]	15	Math and verbal exercises and test practice	Prep school boys
French (1955)[a,c]			
Sample B	4.5	Vocabulary training	Eastern public schoolers
Sample C	15	Math and verbal materials	Eastern public schoolers
FTC (1978)	40	Commercial coaching	New York students from public and private schools
Keefauver (1976)	14	Math and verbal exercises	Private prep schoolers
Lass (1961)[a]	—[d]	Math and verbal exercises	New York high schoolers
Reynolds & Oberman (1987)	63	Math and verbal exercises	Gifted urban students
Zuman A (1988)	27	Math and verbal exercises	New York high schoolers
		Matched comparison studies	
Burke (1986)	50	Advanced reading course	Georgia public schoolers
Coffin A (1987)	18	Computer assisted instruction	Urban public schoolers
Davis (1985)	15	Computer assisted instruction	Florida public schoolers
Frankel (1960)	30	Various kinds of commercial coaching	Bronx High School of Science seniors
Kintisch (1979)	—[d]	Reading instruction	Private prep schoolers
Whitla (1962)[a]	10	Vocabulary, math problems and reading (Reading Institute)[e]	Boston suburb high schoolers

(continued on p. 390)

TABLE 4 (*continued*)

Study	Duration of coaching (hours)	Content of coaching	Type of sample
		Randomized control studies	
Alderman & Powers (1980)[a]			
School A	7	Reading comprehension and analogies	Public schoolers
School B	10	Vocabulary and analogies	Public schoolers
School C	10.5	Reading comprehension and vocabulary	Public schoolers
School D	10	Reading comprehension, analogies and vocabulary	Public schoolers
School E	6	Verbal analogies problems	Public schoolers
School F	5	Verbal analogy problems	Private schoolers in required course
School G	11	Vocabulary and analogies	Private schoolers
School H	45	Reading comprehension, vocabulary and analogies	Private schoolers in elective course
Evans & Pike (1973)[a,c]			
Group QC	21	Quantitative Comparison math items	Public school junior volunteers
Group DS	21	Data Sufficiency math items	Public school junior volunteers
Group RM	21	Regular mathematics items	Public school junior volunteers
Laschewer (1986)	8.9	Computer assisted	New York parochial schoolers
Roberts & Oppenheim (1966)[a]			
Verbal group	7.5	Programmed verbal lessons	Black public school juniors
Math group	7.5	Programmed math lessons	Black public school juniors
Zuman B (1988)	27	Math and verbal exercises	New York minority public schoolers

[a] These nine studies were conducted by or jointly sponsored by the College Board or ETS.
[b] Even though Johnson (1984) used a delayed-treatment control group, the experimental and control group data were combined so a change measure could be calculated.
[c] Experimental groups in these three studies were compared to a common control group.
[d] Duration of coaching was not reported for these studies.
[e] The Reading Institute is a commercial coaching school.

mean-change differences for SAT-V and 26 for SAT-M change. The values of the standardized mean changes are given in Appendix A, Table A-1.

The ranges of values for the standardized mean-change measures for both SAT-M and SAT-V, for coached and control groups, are shown in Table 6. (The control-group means and standard errors shown were the values used to represent nonexistent control-group results.) Both these ranges and the weighted means show that coaching produced larger changes and somewhat more spread in changes than simply retaking the SAT.

No further analysis of the coached-group standardized mean changes is reported because within-study *differences* between coached- and control-group standardized mean changes were analyzed. However, examination of the control-group results provides some information about the use of control-group averages to represent nonexistent control outcomes. The results are also viewed in light of research on gain due to retesting.

Assessment of imputed values. It is impossible to investigate how well the control-group averages represent what might have been found if the uncontrolled studies had had comparison groups. The homogeneity-test values shown in Table 6 represent how well these averages depicted the existing control-group results. Each homogeneity test statistic is distributed as a chi-square statistic with $(k - 1)$ degrees of freedom (k is the number of results) when the studies all share a common population standardized mean change (Becker, 1988). The expected value of this statistic is $(k - 1)$ when the results are perfectly homogeneous.

Homogeneity tests were significant for both SAT-M and SAT-V, indicating that results varied widely across the control samples. The average control-group change

TABLE 5
Characteristics of coaching studies

Characteristic	Category	Number of studies	(%)
Selectivity of sample	Low achievers	6	(12)
	Mix of students	18	(38)
	Selective sample	24	(50)
Voluntariness of sample	Compulsory coaching	3	(6)
	Elective/free coaching	19	(40)
	Voluntary/costly coaching	26	(54)
Content of instruction	Math and verbal	27	(56)
	Math only	4	(8)
	Verbal only	15	(31)
	Neither math nor verbal	2	(4)

Note. Percentages may not sum to 100 because of rounding error.

on SAT-M was 0.16 ($SE = 0.01$) with a homogeneity-test value of $H_T = 152.60$ ($df = 15$, $p < .005$). The average differed significantly from zero, but inconsistency within the control results indicates that some groups may be from control populations in which negative (or conversely very large) changes are expected. The SAT-V average of 0.23 ($SE = 0.01$) also did not adequately represent all control results on SAT-V ($H_T = 143.81$, $df = 27$, $p < .001$).

Although the weighted averages do not seem to represent single *common* values for all control groups, they are reasonable "typical" values. The median standardized mean SAT-M change for the control samples was identical to the weighted average of 0.16 (for SAT-M). The median for SAT-V was 0.16, which is reasonably close to the verbal average of 0.23.

The lack of consistency in the control results indicates that not all control groups performed alike. Differences in either the types of subjects or the activities of the control groups (e.g., wait-list controls versus controls receiving alternative coaching treatments, as in Alderman & Powers, 1980) may have contributed to the variation. Different types of control groups may change to different extents. However, because the values were used to represent *nonexistent* control-group results, the overall weighted average changes were used.

One additional source of information about whether the imputed values are reasonable comes from the literature on score change due to growth and retesting. Bond (1989) proposed a "good guess" of 15 to 20 points for the expected 6-month SAT-M score gain. For the population standard deviation of 100, the standardized mean change of 0.16 would correspond to a gain of 16 points, well within Bond's suggested range. Point gains within Bond's 15- to 20-point range would be obtained for standard deviations ranging between approximately 93 and 125 (with a mean change of 0.16).

Bond noted that expected gains on SAT-V depend on pretest score levels. He suggested 15 points gained for students with SAT-V pretests near 450, with gains nearer to 25 points for students scoring between 500 and 600. Pretest SAT-V scores for coached samples from uncontrolled studies averaged near 500. The SAT-V pretest means were 450 for four studies from Marron (1965) and about 470 for two

TABLE 6
Changes in SAT performance for coached and uncoached samples

Measure	Coached	Uncoached
	SAT-Math	
Range	−0.03 to 1.14	−0.17 to 0.97
Mean	0.47	0.16
(*SE*)	(0.01)	(0.01)
Homogeneity test	—	152.60
Number of samples	26	16
	SAT-Verbal	
Range	−0.26 to 1.08	−0.53 to 0.52
Mean	0.36	0.23
(*SE*)	(0.01)	(0.01)
Homogeneity test	—	143.81
Number of samples	44	28

studies from Pallone (1961). Two studies of college freshmen from Coffman and Parry (1967) had SAT-V pretest means near 555. One lower value was a pretest mean of 363 for Study B from Coffin (1987).

The SAT-V mean-change average of 0.23 corresponds to a 23-point gain on the population score scale, perhaps only slightly more than Bond's values would suggest. Actual SAT-V standard deviations for uncontrolled groups ranged from 61 to 108, corresponding to gains of 14 to 25 points, which are quite close to the range given by Bond.

The imputed values used in this analysis seem like reasonable averages in light of data on gains due to growth and retesting. An alternative approach would be to substitute different predicted mean-change values for each uncontrolled study on the basis of the pretest performance of its coached sample. However, it is not clear how information about implicit imprecision in the substitute values would be incorporated into the analysis.

Initial Analyses of Mean-Change Differences

The first question addressed with the generalized least-squares analysis concerned consistency in the mean-change differences (i.e., the Δ values) across all 70 outcomes from the 48 studies. The regression model included only the grand mean, as shown in the first column of Table 7. The traditional overall homogeneity test (shown as

TABLE 7
Regression analyses for 70 coaching-study outcomes

Predictor	Common Δ	M/V difference	Coaching content	Study design
Grand mean	0.300 (0.014)*	0.255 (0.016)*	−0.388 (0.120)*	0.409 (0.206)*
SAT-M		0.116 (0.018)*	0.117 (0.019)*	0.113 (0.018)*
Control group			−0.026 (0.091)	−0.010 (0.067)
Duration			0.007 (0.001)*	0.006 (0.001)*
Verbal instruction			0.163 (0.073)*	0.202 (0.050)*
Math instruction			−0.041 (0.054)	
Alpha instruction			0.008 (0.029)	
Item practice			0.235 (0.073)*	
Test practice			−0.020 (0.056)	
Test-taking skills			−0.009 (0.044)	
Other activities			−0.034 (0.059)	
Homework			−0.000 (0.054)	
Computer instruction			−0.075 (0.090)	
Wait-list control			0.083 (0.055)	
Alternative control			−0.069 (0.076)	
Year				−0.007 (0.002)*
Publication type				0.004 (0.041)
Use of matching				0.015 (0.063)
Use of randomization				0.238 (0.065)*
ETS sponsorship				−0.198 (0.057)*
Selectivity				−0.033 (0.037)
Voluntariness				−0.004 (0.036)

Note. Standard errors are in parentheses. Asterisks indicate that slope coefficients for predictors differ from zero with $\alpha = .05$.

H_E for the model of a common Δ in Table 8) indicated that results were not consistent across samples and outcomes (H_E = 593.99, df = 69, $p < .005$). The studies did not all appear to share a single Δ value in the population.

Values of the Δ estimates ranged from −0.19 to 0.80 for SAT-M and −0.21 to 0.85 for SAT-V. Several coached groups changed as much as three-fourths of a standard deviation *more* than did their control comparison groups; in other cases, coached groups changed as much as one-fifth of a standard deviation less than the control comparison. Keefauver's (1976) coached group even lost points, whereas the control comparison gained. The sizable homogeneity-test value reflects this wide range of values for the Δ estimates.

The model-significance test (H_R) indicated that the average Δ value, estimated to be 0.30, differed from zero ($p < .001$). Thus, although not all coaching studies can be considered to show this degree of change, on average coached groups in this review gained three-tenths of a standard deviation more than did uncoached groups.

The next regression equation predicted mean-change differences from 21 predictors. The omnibus test for the combined effects of all predictors (excluding the grand mean) was highly significant (H_R = 430.88, df = 21), and the model explained 72% of the variation in mean-change differences. However, the test for remaining variability was also significant (H_E = 163.06, df = 48, $p < .005$); thus, even this overall regression did not fully explain the patterns of study results.

TABLE 8
Model tests for analyses of coaching-study outcomes

Model	Model significance		Model specification		Percentage of variance explained
	H_R	(df)	H_E	(df)	
		All results			
Common Δ	465.41	(1)	593.99	(69)	
Math/verbal differences	40.75	(1)	553.19	(68)	6.9
Coaching content	420.92	(13)	173.03	(56)	70.9
Study design	417.32	(11)	176.63	(58)	70.3
		Published results			
Common Δ	35.97	(1)	100.34	(24)	
Math/verbal differences	2.11	(1)	98.22	(23)	2.1
Coaching content	75.46	(11)	24.88	(15)[a]	75.2
Study design	70.82	(8)	29.52	(16)[b]	70.6
		Unpublished results			
Common Δ	551.29	(1)	394.38	(44)	
Math/verbal differences	41.89	(1)	352.50	(43)	10.6
Coaching content	278.11	(12)	116.27	(31)	70.5
Study design	274.29	(10)	120.09	(34)	69.5

Note. All model significance and specification tests are significant at the $\alpha = .005$ level, except as noted below.

[a] This model specification test was significant at the $\alpha = .10$ level.

[b] This model specification test was significant at the $\alpha = .025$ level.

Approximate z tests for the significance of individual predictors showed seven predictors to be significant. The publication date, duration, ETS sponsorship, presence of coaching for SAT-M, instruction in test-taking skills, assignment of homework or other out-of-class activities, and the difference between math and verbal outcomes all were significant.

Inspection of the correlation matrix among the slopes in the 21 predictor model revealed multicollinearity problems among several of the predictors. Many of the zero-order correlations among the predictors differed significantly from zero. (Values of the predictor variables are given in Appendix B, Table A-2.)

Several characteristics appeared confounded with the use of a comparison-group design, including duration of coaching (comparison studies were briefer), ETS sponsorship (most ETS studies used comparison groups), and the degree to which subjects had volunteered for the study (comparison studies more often had volunteer subjects). Characteristics of comparison groups (e.g., use of matching or randomization, type of control group) were correlated with the indicator for comparison versus uncontrolled studies. Also, longer studies tended to use more selective and less voluntary samples and typically were not conducted by ETS. Finally, more recent studies tended to be randomized.

Three specific reduced models were estimated, and results of these analyses are given in Tables 7 and 8. Each model revealed certain relationships of predictors to the degree of change in studies of SAT coaching. However, no model fully explained the patterns of study outcomes (as indicated by significant model-specification tests in Table 8). The results only partially explained the differences in the efficacy of coaching and the influences of study and sample characteristics on coaching outcomes.

The second model in Tables 7 and 8 examined whether the effects of coaching differ for SAT-M and SAT-V, whereas other characteristics of studies were allowed to vary. Advantages for coached groups averaged 0.12 standard deviations larger for SAT-M than for SAT-V. The slope value was quite stable in the context of other models as well.[9] That is, the size of the SAT-M versus SAT-V difference did not change much when other variables were controlled. However, the model of different coaching effects for SAT-M and SAT-V explained only 7% of the variation in study outcomes.

The third model in Table 7 shows the contributions of characteristics of the coaching interventions to the prediction of differences in change. Three instructional variables were significant—the duration of the program, the presence of verbal instruction, and the presence of item practice as a coaching activity. Longer interventions were more effective, although the effect was rather small, with coached-group advantages of only 0.07 expected for each 10-hour period of coaching. Interventions with instruction on verbal skills showed coached-group advantages; however, only 6 of 40 samples had *not* provided verbal instruction.

Instruction aimed at increasing what Bond (1989) described as alpha abilities contributed positively but not significantly to the coaching effects in these studies. However, studies in which the coaching intervention included practice and instructions on answering particular items (and item types) did show significant advantages (with the effects of other factors held constant). Across all studies, this factor apparently had a greater impact than practice in taking complete exams or instruction in general test-taking skills.

The fourth model in Table 7 estimated the effects of various aspects of the design of coaching studies. Again duration, the SAT-M versus SAT-V effect, and the effects of verbal instruction were significant, as were three design characteristics.

The negative slope for year of publication indicated that coaching effects are decreasing over time in this literature. Also, the indicator of ETS sponsorship of the study was significant and negative. Studies sponsored by ETS showed smaller coaching advantages, with other factors held constant.[10]

Larger coached-group advantages for samples from randomized studies are reflected by the positive slope for randomization. Randomized studies showed coached-group advantages averaging 0.76 standard deviations, with other factors held constant. However, only one of 18 estimates from randomized studies (from Zuman, 1988) was actually larger than 0.30. This suggests that although randomized studies showed smaller raw (absolute) effects, they had other characteristics that were more generally associated with smaller effects (such as simply the presence of a control group). Once adjustments for these other characteristics were made, the randomized studies had larger effects than would have been expected. Presence and type of control group did not contribute to prediction of the coaching effects, although all three predictors showed negative slopes, indicating larger effects for uncontrolled studies.

Table 8 shows the overall model-significance and specification tests for the four models outlined in Table 7. The more complex models explained more of the variation in coaching effects; these models explained roughly 70% of variation in the study outcomes.

Analyses of Published and Unpublished Studies

Earlier reviews of the coaching literature had indicated that results of uncontrolled studies were more dispersed than those of comparison studies. The Δ estimates for comparison and uncontrolled studies were analyzed separately, and initial homogeneity tests supported the finding of slightly more variation among uncontrolled results ($H_T = 137.09$, $df = 14$) than among comparison studies ($H_T = 147.32$, $df = 38$). However, exploratory regression modeling failed to produce adequate explanatory models for the separate sets of results. These results are not detailed here.

A second breakdown of results suggested by previous meta-analyses considers published and unpublished studies. Several reviews in other domains have found more variable results, and sometimes differing results, according to the source of the study (see, e.g., Smith, 1980). Kulik et al. (1984) found only minimal differences in average coaching effects (with smaller effects in unpublished studies). They also reported slightly more variation among the unpublished results.

Regression modeling of published and unpublished studies provided some interesting and adequately fitting explanatory models for coaching effects. The analyses involved fitting separate regression models for the two subsets of coaching-study outcomes. Thus, the tests and analyses reported in this section ignore some of the intercorrelations between the comparison studies and the proxy control-group results in the uncontrolled studies. The following analyses of results of the published and unpublished studies are not strictly independent, and they are not independent of the analyses reported above.

Published studies. The set of published studies consisted of all results reported initially in academic journal articles. For example, the study by Dear (1958) was

therefore considered unpublished even though some details of the research have appeared in journal articles (i.e., in French & Dear, 1959). Conference presentations, dissertations, unpublished reports, and government documents were considered unpublished. Some authors consider publication status to be a proxy for study quality, because many unpublished documents have not undergone the review process typical of most academic journals.

Fitting the model of a single Δ value (the first model in Table 9) to the 25 results from published studies produced an overall homogeneity test value of 100.34, which was highly significant ($p < .005$) based on a chi-square distribution with 24 degrees of freedom. (Overall model tests are in Table 8.) However, the overall homogeneity-test value for the published studies (i.e., for 36% of the full set of results) was less than 20% of the size of the overall homogeneity value for all studies. The simplest possible model, of a population mean-change difference of 0.14 standard-deviation units, showed that coaching effects were generally smaller in the published literature.

A model based on all coded predictors[11] showed severe multicollinearity. Models representing math versus verbal differences, coaching-content, and study-design effects were estimated as for the full set of results, with slight modifications in the models due to problems of linear dependencies among the predictors, as described below.

The SAT-M versus SAT-V difference was not as pronounced for published research as in the overall set of studies, but slightly stronger effects of coaching on math outcomes can be seen across all models in Table 9.

TABLE 9
Regression analyses for 25 published coaching-study outcomes

Predictor	Common Δ	M/V difference	Coaching content[a]	Study design
Grand mean	0.145 (0.024)*	0.132 (0.026)*	−0.484 (0.141)*	2.203 (0.759)*
SAT-M		0.044 (0.030)	0.077 (0.032)*	0.079 (0.032)*
Control group				−0.315 (0.178)
Duration			0.002 (0.002)	0.001 (0.003)
Verbal instruction			0.013 (0.114)	0.213 (0.070)*
Alpha instruction			−0.362 (0.122)*	
Item practice			0.905 (0.182)*	
Test practice			0.181 (0.113)	
Test-taking skills			0.569 (0.197)*	
Homework			−0.568 (0.175)*	
Year				−0.023 (0.007)*
Use of randomization				0.762 (0.193)*
ETS sponsorship				−0.616 (0.198)*
Selectivity				−0.114 (0.105)
Voluntariness				−0.084 (0.082)

Note. Standard errors are in parentheses. Asterisks indicate that slope coefficients for predictors differ from zero with $\alpha = .05$.

[a] This model shows multicollinearity between predictors for alpha instruction, item practice, and homework.

The coaching-effects and study-design models estimated for the published studies differed from the model used for all studies because dependencies among variables that did not appear in the full set of results arose in the reduced set of cases. A model including all predictors shown in the coaching-effects model in Table 7 (except publication type) could not be estimated. Many dependencies appeared to involve the uncontrolled- versus comparison-study dummy variable, which was eliminated from subsequent analyses. Other predictors, such as the indicators of matching and of other activities, were eliminated because of little variation in predictor values. The models are not strictly comparable to the overall models (or ones estimated for unpublished studies), because different sets of predictors have been included.

Results of the coaching-content model for published studies differed from the model for all studies. The SAT-M versus SAT-V difference was significant, although smaller for the published results, and item practice also showed a strong positive contribution. Duration and presence of verbal instruction were nonsignificant, although their slopes were positive (as for the model based on all results). Additionally, practice completing sample tests had a large contribution, whereas both instruction on alpha abilities and use of homework and outside assignments showed strong negative effects on outcomes. Thus, after other features of the coaching instruction have been held constant, studies that presented alpha instruction and that assigned homework showed smaller coaching effects.

Multicollinearity among the predictors in this model was indicated, however, by high correlations among the slopes. The correlations of the slope for test-taking skill instruction with slopes for alpha instruction and homework were −.89 and

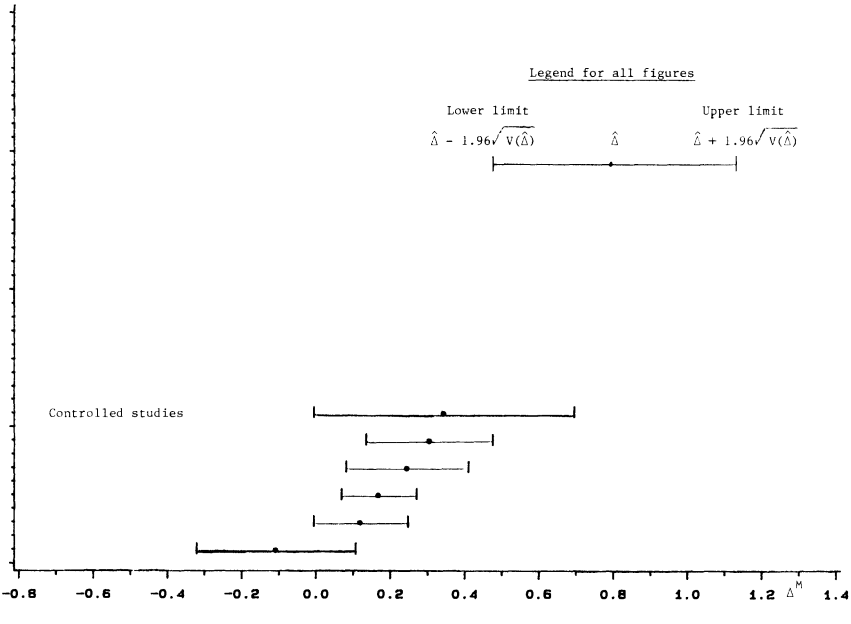

FIGURE 1. *Ninety-five percent confidence intervals for SAT-M coaching effects for published studies*

−.82, respectively. Removing the predictor for instruction in test-taking skills from the model caused both alpha instruction and homework to have nonsignificant slopes, whereas duration then had a significant positive slope.[12] The extreme multicollinearity in the coaching-content model suggests that it might be unwise to use this model for predicting the effectiveness of new coaching interventions.

The effects of study design on coaching effects in the published studies were similar to those found for the full set of results, even though the models again included slightly different sets of predictors. All slopes had the same signs, and many were of the same order of magnitude as for the full set of results. One exception was the duration of coaching, which was positive but not significant in the published literature. Again, use of randomization had a significant positive contribution, which was countered somewhat by a negative slope for presence of any control group. Essentially, these results suggest that published studies with comparison groups did not show smaller coaching effects than uncontrolled published studies.[13]

The fit of the models for published studies was better than for the full set of results. The model-specification test for the coaching-content model was not significant at the .05 level. Additionally, standardized residuals[14] from the coaching-content model were all small.

Figures 1 and 2 show 95% confidence intervals for the SAT-M and SAT-V Δ estimates obtained from published studies. Results from controlled and uncontrolled studies are separated. Within these groups, effects are arranged in order of magnitude, and coaching effects are denoted by the · symbol. Figure 2 shows that the verbal results from the two uncontrolled published studies vary much more

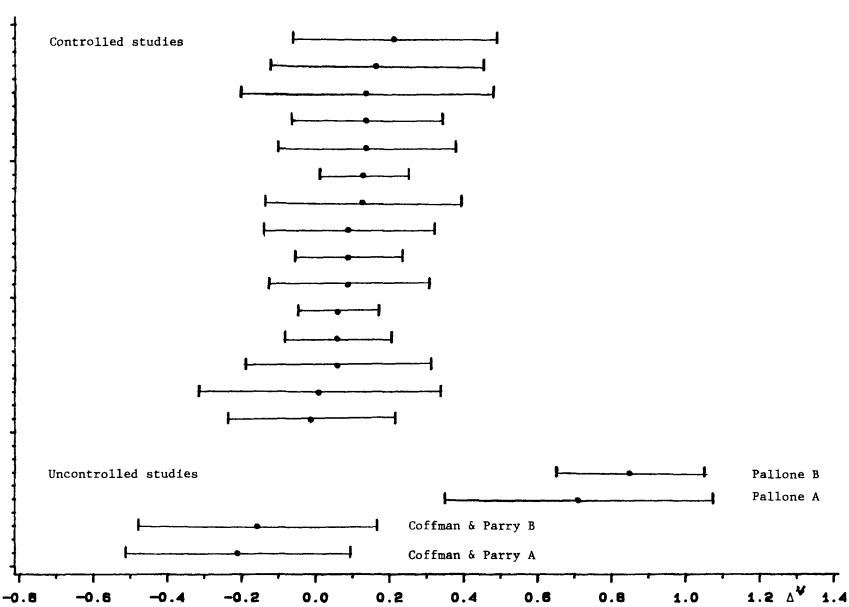

FIGURE 2. *Ninety-five percent confidence intervals for SAT-V coaching effects from published studies*

than those from controlled published studies. In fact, when the results from Coffman and Parry (1967) and Pallone (1961) are deleted, the remaining 21 effects can be adequately modeled by a regression including only the predictor for SAT-M versus SAT-V differences. Specifically, the model is

$$\hat{\Delta} = 0.088 + 0.069 \text{ (SAT-M)},$$

which produces significant mean effects of 0.09 standard deviation units for SAT-V outcomes and 0.16 for SAT-M outcomes. The math-verbal difference is significant ($H_R = 5.09$, $df = 1$), and the residual error is nonsignificant ($H_E = 27.65$, $df = 19$, $p > .05$). Thus, by eliminating only 4 (16%) of 25 results, a very simple "fixed-effects" model can be estimated that adequately describes all variation in outcomes beyond what would be expected due to sampling error.

Unpublished studies. The 45 results from unpublished studies were highly inconsistent, with a homogeneity-test value of 394.38 ($df = 44$, $p < .001$; see Table 8). This value, based on 64% of the results in the review, was about 66% of the size of the homogeneity value based on all 70 results. (Because of the covariances between the imputed control results and the comparison-study results, the subgroup homogeneity values for published and unpublished studies do not sum to the total for all studies of 593.99). Figures 3 and 4 show the math and verbal outcomes for unpublished studies.

The estimated common Δ value for the unpublished studies was 0.37, more than twice as large as the average for the unpublished studies. Table 10 shows this highly significant value. The unpublished papers and reports in this collection showed

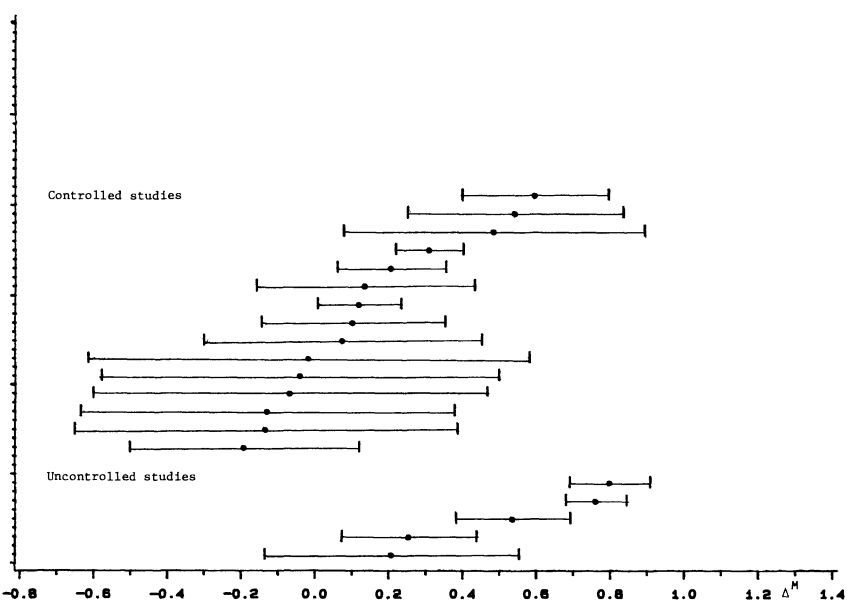

FIGURE 3. *Ninety-five percent confidence intervals for SAT-M coaching effects from unpublished studies*

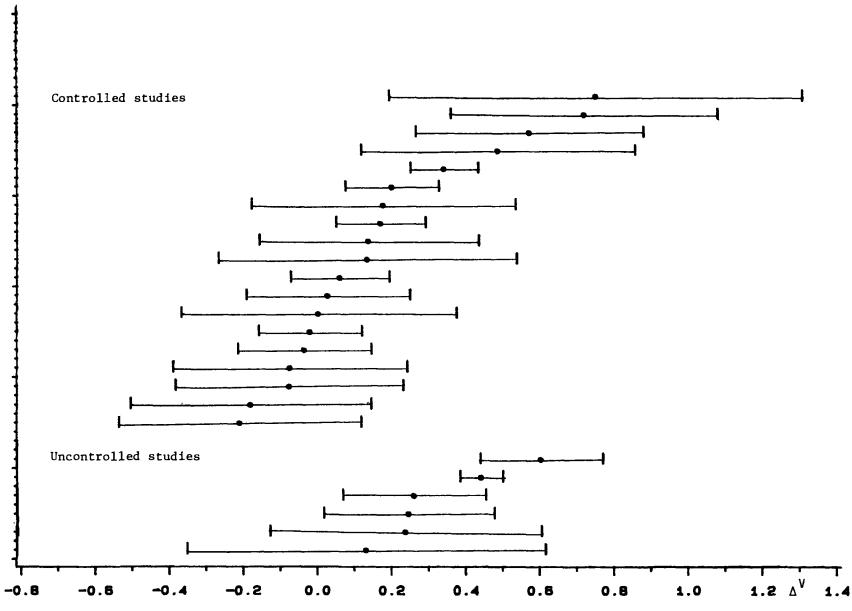

FIGURE 4. *Ninety-five percent confidence intervals for SAT-V coaching effects from unpublished studies*

stronger coaching effects than studies found in academic journals, when other study features were allowed to vary.

Regression analyses were computed for the unpublished results, and summaries of the models are reported in Tables 8 and 10. The mathematical versus verbal score difference was larger for the unpublished results than for published studies; the "SAT-M" slope values were approximately twice those from Table 9.

Among the predictors representing the nature of the coaching intervention, only the presence of test practice appeared to produce significant advantages for coached groups in these unpublished studies.

Treatment duration showed a positive and significant contribution, as did the indicator of verbal instruction (each in the study-design model). The effects of design features were similar to those found for the published sets of results, in spite of differences in the set of modeled predictors. Year of publication and ETS sponsorship again had negative, but nonsignificant, associations with coaching effects. The coefficient for matching was large, but it represented only one study. Study A from Coffin (1987) was the only unpublished study using matching, and its verbal $\hat{\Delta}$ value was the second largest value across all studies.

A final regression model was created by using only the intercept and the coefficients that were significant in the coaching-content and study-design models. These five predictors produced a regression equation that explained 67% of the variation in outcomes, but only the effects of duration and SAT-M versus SAT-V were significant. Apparently, interactions among the included and the omitted predictors are contributing to multicollinearity in the models shown in Table 10.

TABLE 10
Regression analyses for 45 unpublished coaching-study outcomes

Predictor	Common Δ	M/V difference	Coaching content	Study design
Grand mean	0.373 (0.016)*	0.312 (0.018)*	0.243 (0.380)	0.043 (0.649)
SAT-M		0.143 (0.022)*	0.149 (0.023)*	0.146 (0.023)*
Control group			−0.101 (0.169)	0.038 (0.092)
Duration			0.005 (0.005)	0.006 (0.002)*
Verbal instruction			−0.009 (0.132)	0.222 (0.083)*
Math instruction			−0.240 (0.123)	
Alpha instruction			0.053 (0.047)	
Item practice			−0.089 (0.200)	
Test practice			0.301 (0.153)*	
Test-taking skills			−0.029 (0.082)	
Other activities			0.013 (0.091)	
Homework			−0.114 (0.115)	
Computer instruction			0.112 (0.148)	
Wait-list control			0.086 (0.076)	
Alternative control			−0.178 (0.159)	
Year				−0.004 (0.008)
Use of matching				0.251 (0.114)*
Use of randomization				0.036 (0.106)
ETS sponsorship				−0.038 (0.176)
Selectivity				−0.075 (0.052)
Voluntariness				0.058 (0.068)

Note. Standard errors are in parentheses. Asterisks indicate that slope coefficients for predictors differ from zero with $\alpha = .05$.

The overall fit of the regression models for the results of unpublished studies was nearly as good as for published study results. However, again all model-specification tests were significant ($p < .005$). Standardized residuals for both models for unpublished results showed specific results were consistently poorly predicted. Eleven results had large residuals in the coaching-effects and study-design models, and six of those results fit poorly in both models. Four of the residuals were from Marron (1965), one was from Reynolds and Oberman (1987), and one was from Zuman (1988, Sample A). The largest discrepancy between a predicted and an actual value was for Coffin's (1987) matched samples. The coached group showed a larger than expected advantage ($\hat{\Delta} = 0.75$) on SAT-V, which was due mainly to an extreme loss of points for the control group. Coffin reported serious morale and motivation problems in the small group (which had to walk some distance from school to participate in the alternative-treatment activity of using the College Explorer computer program). The predicted Δ value was −0.16 (based on the coaching-content model).

The coaching-content and study-design models for unpublished results produced significant standardized residuals for 10 and 7 results, respectively. That is, 22% and 15% of the results were poorly predicted by these models. Only one result, or 4% of the published results, had a significant residual. By the criterion of number of large regression residuals, the models for published results explained more of the variation in study outcomes than the models for unpublished results.

Summary of the separate analyses. No single model based on the initial predictors (plus the grand mean) adequately described the results of the 70 coaching-study outcomes taken together. Results were separated into those from published and unpublished studies, and the elimination of results from two published but uncontrolled studies produced homogeneous results among published studies. Models for both sets of studies explained major portions of the variation in study outcomes; up to 76% of the variation in published results was explained. The effects of significant predictors were fairly similar in models for the two sets of results, differing for the most part only in magnitude.

Models for both coaching-content and design effects accounted for large portions of the variation in results for the two subsets of studies. Using the criterion of "variance accounted for," both sets of models appear to provide strong explanations of coaching effects. For each set of studies, the coaching-effect and study-design models provide alternative explanations of the study results.

Limitations

The analyses described above are limited largely by the nature of the data under review. In particular, the results depend on the validity of the measures of the predictor variables and outcomes themselves. On the surface, the outcome would appear to be fairly well understood. The predictors are somewhat more problematic.

The SAT is a standardized test given several times each year. The Preliminary Scholastic Aptitude Test, which parallels the form of the SAT, is also technically sound. All of the measures used in studies of SAT coaching should be equated with the SAT, and, if they are truly equated, analyses can even be conducted in terms of raw points gained. As discussed above, it is not clear that the measures used in the coaching literature are always comparable to the "real" SAT. However, studies using nonstandard SATs did not appear to be responsible for unexplained variation in study outcomes (i.e., they typically "fit" with the estimated models). Thus, we have some evidence that the differences in outcome measures across these studies do not constitute a serious limitation on the conclusions.

The quality of the measures of predictors of coaching effectiveness also bears examination. Many values of important predictors could not be determined from the reports of coaching studies. Duration was *not* explicitly reported in 9 of 23 reports. Only Keefauver's (1976) dissertation gave a detailed account of how time was spent *during* the coaching intervention. Consequently, the measures of instructional activities and emphases (e.g., alpha instruction) in this review are, at best, crude indicators of the content of the coaching interventions.

Most troublesome was the lack of information about the coaching interventions. Three reports gave no information about materials used, and four more described exercises only very generally. Seven reports identified commercial products that were used, but little information was given about *how* they were used. In some samples (e.g., the FTC study) subjects were drawn from a number of different coaching programs, which may have differed in unknown ways. Eleven of 19 reports on comparison studies (almost 60%) gave no information about the activities of their control subjects. This is a serious problem in a literature where group comparisons are of paramount importance.

Orwin and Cordray's (1985) study of deficient reporting in meta-analysis showed that the role of study characteristics can be obscured when information has been

poorly reported. In their study, analyses in which measures of study quality were disattenuated suggested that study quality had a larger impact on study results as compared to analyses based on raw coded study-quality data. The consequences of poorly reported data in the literature on coaching effects cannot be predicted.

Discussion

These analyses provide answers to the questions posed at the beginning of the review, as well as a perspective on research on SAT coaching and implications for future research. The discussion is organized around issues raised in past reviews of the coaching literature.

Coaching Content

The first question posed in this review concerned the relative contributions to the results of characteristics of the subjects, the coaching interventions, and studies themselves. Individual and combined effects of several kinds of coaching instruction were examined empirically.

Item practice and instruction in test-taking skills had a significant impact on results for published studies, and practice on sample tests was related to coaching advantages for unpublished results. As Bond (1989) noted, most standardized tests are constructed so that the impact of coaching on beta abilities should be small. That is, a principle of good test development is to construct items so that examinees' test wiseness and familiarity with item formats have a minimal relationship with item performance. However, increasing an examinee's familiarity with novel item types (such as the SAT-M's data-sufficiency or quantitative-comparison items) may well enable him or her to improve SAT performance considerably.

The absence of positive effects for the presence of instruction of alpha abilities and the use of homework is puzzling, given the claims of the developers of the SAT. Content-relevant instruction, which likely bears the greatest resemblance to more formal and ongoing schooling, should increase coached-group advantages. However, given the minimal level of reporting about the actual coaching interventions, the coding of instruction for alpha abilities may have been inadequate. (This issue cannot be explored empirically with the present data.)

Similarly, no information about homework was available for 14 of 48 studies (nearly 30%). In the analyses these studies were treated as if no homework had been assigned; if homework was actually assigned, but not reported in these studies, the results for the homework predictor may be misleading.

Heavily content-oriented instruction is completely in accord with what others have called "proper preparation" for standardized testing (Mehrens & Kaminski, 1988). Mehrens and Kaminski (1988) suggested that it may be somewhat less ethical to provide instruction organized around objectives drawn by looking at, for example, the objectives or content of a particular instrument (the kind of instruction I have termed *item practice*). This is a particular concern in the published literature, in which item practice contributed strongly to coaching outcomes.

Program Duration

In half of the analyses reported above, duration of coaching was only minimally related to study outcomes. Its effect was weak, and the impact was generally not

significant after instructional and design differences had been controlled. Duration did not relate to coaching effectiveness after experimental-design factors were controlled in published studies, but had a slight positive relationship to coaching effects in the unpublished literature. In part, this was due to restriction of range in the duration variable (especially within the subset of published results). However, duration was also collinear with a number of other predictors in models for the published results, suggesting that other significant variables (e.g., test-taking skills) may have been representing the same effects as duration. For published *comparison* studies, duration was unrelated to coaching effects, even when other characteristics of programs were not accounted for.

Nonetheless, the results are not at odds with Messick and Jungeblut's (1981) conclusions. Messick and Jungeblut had used duration as the only predictor in their regression analyses,[15] and in their data, as well as here, duration was confounded with design features. This confounding increased the apparent effect of duration in their analysis, and decreased it here.

The unpublished studies suggest that students should only expect to make great gains in performance by spending considerable time in coaching. A coached-group advantage of only one to six points on the population score scale is expected for every 10 hours of coaching. This agrees moderately well with Messick and Jungeblut's (1981) "threshold" of 3 hours of coaching needed to produce any SAT gain beyond that expected upon retesting.

Study Quality

Study design was represented in these analyses by indicators of presence and type of comparison group, as well as the mechanism used to group subjects (e.g., randomization). Few generalizations can easily be made about the role of study quality in the coaching literature. For the published literature, one model of outcomes showed larger coaching effects for randomized studies (when other factors were held constant). Statistical design considerations suggest that randomized comparison studies have more internal validity than studies using the single-sample pretest-posttest design (see, e.g., Campbell & Stanley, 1963); thus, this difference warrants some attention. The result agrees with similar findings by Kulik et al. (1984), even though it is based on results from only two sources (including 14 randomized-study results).

A significant advantage for matched unpublished studies can be attributed to the results of a single study using matching. Across all studies, the type of comparison group used (wait-list versus alternative-treatment control) did not relate to the size of the coaching effects.[16]

Study quality is presumed to be one factor that affects whether or not a study is published. Sixteen percent of published results in the review were uncontrolled studies versus 24% of the unpublished results. When all published results from comparison studies were analyzed, findings were consistent with a very simple model of coaching effects. Advantages for coached groups of 9 SAT-V points and 16 SAT-M points (on the population score scale) were predicted. If we consider these studies to provide the most rigorous evaluation of coaching's potential, we must expect only modest gains from *any* coaching intervention. This is the clearest finding of the synthesis.

The current analysis also addresses a question raised by the existence of different conclusions in past reviews. Messick and Jungeblut (1981) found duration effects when duration was the only predictor of coaching analyzed. Likewise, DerSimonian and Laird (1983) found design effects, but did not explore the role of duration. The present analyses suggest that duration and design effects are not substantial enough to persist when considered together.

Two other study characteristics of note were the publication date and ETS sponsorship. Both of these had negative effects on study outcomes, although the extent of change over time suggested by the results was small. Coached-group advantages appear to be shrinking by no more than 2 points per year on the population SAT scale where the standard deviation is 100. Differences between studies sponsored by ETS and others were substantial, especially in the published literature.

Several factors may be contributing to decreases over time in coaching effects. More test-preparation materials have become increasingly available from both test developers (Cole, 1982) and commercial vendors (e.g., Staples, 1985). Any students willing to expend their own time can now find many materials for SAT preparation in public libraries and at relatively low cost in bookstores. Powers (1982) found that approximately half of a national sample of SAT examinees had prepared for the test using test-preparation booklets, whereas over three-fourths of the sample had completed sample items in the test-familiarization booklet *About the SAT*. Other influential factors may include new laws on test disclosure (although evidence suggests that disclosure laws have little effect, according to Bond, 1989), and more awareness of the impact of testing and sensitivity to testing-related concerns on the part of the general public (e.g., Lewis, 1989). This could lead to a higher general level of self-study in both alpha and beta abilities, increasing comparison-group scores and thereby decreasing treatment effects.

Studies sponsored by ETS were associated with smaller coaching advantages than other studies, a difference that was significant in the published literature. Critics have disparaged ETS and the College Board for maintaining the position that coaching for the SAT is ineffective (Owen, 1985; Slack & Porter, 1980). The present finding is evidence that the ETS stance is based on different evidence than is available from the full collection of coaching studies. Published studies sponsored by ETS show much smaller coaching effects even when other design features are held constant.

Yet clearly, the base of evidence that ETS has long cited consists primarily of studies designed with attention to statistical control. The ETS-sponsored studies in this collection, with the exception of the Coffman and Parry (1967) and Marron (1965), were comparison studies. Although the external validity of some of their studies is questionable (e.g., the use of a sample of college freshmen by Coffman and Parry is quite problematic), it is difficult to fault ETS for relying on accepted principles of experimental design in downplaying the results of uncontrolled studies.

Another study-design concern revealed in this study was the extensive interrelatedness of coaching characteristics. Past research, zero-order correlations among the predictors, and correlations among regression weights in this analysis indicated that some characteristics were highly interrelated and thus confounded. Notably, comparison and uncontrolled studies differed not only in their experimental designs, but also in their duration, the degree of voluntariness of their samples, and their

sponsorship by ETS. Thus, differences between comparison and uncontrolled samples may have resulted from a number of influences. Clearly, further study (i.e., further primary research) would be necessary to fully disentangle the effects of duration, date of publication, and experimental design.

Motivation and Selectivity

A central issue in the coaching literature revolves around the role of student self-selection into coaching. Studies of commercial coaching (e.g., FTC, 1978, 1979; Zuman, 1988) have been critiqued intensely on that issue. As Powers noted in a broader study of methods of SAT preparation, "random assignment to test preparation methods seems especially desirable for studies of the effects of special preparation, in which effects may be quite small relative to the potential effects of self-selection factors" (1982, p. 18).

In the present analyses self-selection effects were represented by the "voluntariness" variable. The impact of this predictor was essentially nil. What does "voluntariness" of these samples represent? Samples were given high values on voluntariness when students had sought and paid for commercial coaching (e.g., Study A from Zuman, 1988) and when they had participated in coaching as an extracurricular activity. Students receiving coaching as a school elective were labeled as moderate on voluntariness. Motivation to improve scores would likely be high in both groups, but students who have paid for coaching may also differ in other ways. Specifically, they have the financial resources to spend on coaching.

Possibly, coding the voluntariness variable as a dichotomy, to differentiate students who have paid for coaching from those who obtained free extracurricular coaching, would have produced different results. Kulik et al. (1984) compared commercial and school-based programs, but found no differences between them. With the present coding scheme, more than half of the samples were given high voluntariness values. However, with the current conception of voluntariness, self-selection effects do not appear strong in the coaching studies after other treatment and study characteristics have been held constant. Essentially, this indicates that coaching will be equally effective for students who have volunteered and paid for coaching and those for whom coaching is compulsory.

Differential Effects of Coaching on SAT-M and SAT-V

The second overall question of the review was whether the effect of coaching was the same for SAT Verbal and Mathematical performance. All regression models indicated that the effect of coaching was stronger for math than for verbal outcomes. Although the magnitude of the difference was smaller for published studies, it was nonetheless significant in most models. Estimates of the size of the difference between math and verbal effects ranged from a minimal 0.04 standard-deviation units ($SE = 0.03$) for published studies to 0.15 ($SE = 0.02$) for unpublished results.

This SAT-M advantage is consistent with the idea that SAT-M performance (and mathematics performance more generally) is more easily coached than verbal performance. Becoming familiar with algorithms and overlearning a few generally applicable mathematical formulas may have a large impact on student performance, whereas vocabulary review may need to be more extensive to have an equivalent impact. Also, familiarization with some of the more unusual SAT-M item types

such as Quantitative-Comparison or the obsolete Data-Sufficiency items may have a larger impact (e.g., compared to practice on verbal-analogy items) because of the novel forms of these items.

Assessment of Adequacy of Regression Models

The third and fourth questions addressed the issue of determining a well-specified model for the coaching-study effects. Initial analyses showed that it was impossible to fit a single model that would completely explain variation across all results. Two regression models based on published results showed nonsignificant residual variance and had no individual results that were poorly predicted. However, models of results of unpublished studies always showed significant unexplained variability and had numerous "misfitting" results. This is consistent with other findings that excessive variation in study results is associated with poorer study quality in other domains (e.g., Hedges, 1986). Published studies, having undergone the process of review and revision, are often believed to be somewhat higher in quality than unpublished research.

None of the results of published studies appeared unreasonable (misfitting) in light of the proposed explanatory models. Unpublished results from studies by Coffin (1987), Keefauver (1976), Marron (1965), and Reynolds and Oberman (1987), as well as from the FTC study (Rock, 1980), were not well-predicted by the models for unpublished studies. The only shared feature that seemed to contribute to their poor fit was the fact that many had used large samples. Only the residuals from Reynolds and Oberman (1987) were based on a nonstandard SAT measure (the PSAT). The fact that other studies had used the PSAT, and that the residuals for the math and verbal results from this study are opposite in sign, suggests that a systematic bias related to the measure used is unlikely.

Implications for Coaching and Research on Coaching

Across all studies, the magnitudes of coaching effects depend on many factors. Some positive instructional effects were accompanied by the influence of design factors. Interrelationships among the characteristics of studies were complex, and to some extent precluded accurate assessment of the importance of all predictors to coaching outcomes.

This review has validated several past concerns. Future research must attend to the unconfounding of subject and study characteristics. In particular, additional controlled studies of longer-term coaching interventions would be useful. More uncontrolled, poorly designed studies of vaguely described coaching programs will only further muddy this currently murky literature.

Furthermore, primary researchers investigating interventions purported to be "coaching" should provide detailed descriptions of the activities involved, materials used, and time spent in instruction. It is difficult to make specific recommendations about what constitutes "effective coaching" when it is not at all clear what constitutes "coaching." The state of the literature and of the research on coaching interventions hinders efforts to make good recommendations concerning policy and practice. Until more well-described, well-designed studies have examined a broader range of coaching interventions, a clear understanding of the contributions of all facets of the coaching process will be unattainable.

Notes

[1] Raudenbush, Becker, and Kalaian (1987) described the use of this approach with data represented by Glass's effect size; the methodology generalizes easily for use with standardized mean changes.

[2] The regression analyses were based on standard least-squares estimation procedures. Error due to sampling variation and differences in sample sizes was not accounted for in these models.

[3] Because their review included aptitude tests with results reported on scales other than the 200–800-point SAT scale, these reviewers represented study outcomes in terms of Glass's (1976) effect size, or standardized mean difference. Glass's effect size contrasts means of coached and uncoached students; thus, results of uncontrolled studies were not included in this review.

[4] For example, students in Roberts and Oppenheim's group were Black high schoolers from Tennessee, whose pretest scores were expected to be "equivalent to about 300 on the SAT, a level of performance appreciably below that found in previous coaching studies" (1966, p. 2). Conversely, all of Marron's (1965) subjects were from strictly college preparatory schools.

[5] This problem arises in other literatures (e.g., the research on psychotherapy outcome studies). Typically, reviewers have had to ignore such dependencies between study outcomes, although that approach is not optimal.

[6] The seven samples from this study were given the same coded values; thus, one error was counted seven times. If the seven samples are only counted once, the reliability of test practice increases to 81%.

[7] Although all subjects in Marron (1965) took both the math and verbal subtests of the SAT, subjects were grouped differently in order to report the results of SAT-M and SAT-V. Because of the complexity of the subgroup arrangements, the intercorrelations between the math and verbal standardized-mean-change differences for this study were considered equal to zero.

[8] Standard deviations pooled across studies were smaller for the underachieving and highly selective (e.g., private school) samples.

[9] The SAT-M slope was estimated to be between 0.10 and 0.13, regardless of which predictors were included in the equations. The value was stable in more than 15 different regression models.

[10] The ETS-sponsorship indicator variable was confounded with the predictors for use of randomization and study duration (which both showed positive slopes). Omitting the "randomization" predictor produced a smaller, nonsignificant negative slope for ETS sponsorship.

[11] Not all predictors could be used in the analyses of subsets of results. For instance, the predictor for use of computerized instruction was identically zero for published studies.

[12] The fit of this latter model was not as good as for the tabled model; it explained only 67.4% of the variation in the study outcomes.

[13] Only two published studies were uncontrolled. Pallone's (1961) study had produced the largest coaching effects, whereas Coffman and Parry's (1967) samples showed nearly the smallest effects of the uncontrolled studies. On average, they were not larger than the comparison-study effects.

[14] Standardized regression residuals in meta-analysis have approximate standard normal distributions (see, e.g., Hedges & Olkin, 1985). It is typical to consider standardized residuals larger than ± 1.96 to be unusual.

[15] Messick and Jungeblut had used log-time rather than the actual length of the coaching program in hours. Logarithmically transforming the duration values in this data set provided a poorer fit than using actual program duration.

[16] For the published studies the two type-of-control-group predictors were both significant and negative in models not presented in Table 9.

APPENDIX A
TABLE A-1
Results of coaching studies

Sample	n^C	n^U	d^{CM}	d^{UM}	d^{CV}	d^{UV}
			Uncontrolled studies			
Coffin B	10	.	0.40	.	0.51	.
Coffman & Parry A	10	.	.	.	0.02	.
Coffman & Parry B	9	.	.	.	0.08	.
Johnson	38	.	0.42	.	0.50	.
Marron 1	83	.	.	.	0.84	.
Marron 2	600	.	.	.	0.67	.
Marron 3	5	.	.	.	0.45	.
Marron 4	26	.	.	.	0.49	.
Marron 5	232	.	0.96	.	.	.
Marron 6	405	.	0.92	.	.	.
Marron 7	78	.	0.70	.	.	.
Pallone A	20	.	.	.	0.98	.
Pallone B	80	.	.	.	1.09	.
			Nonequivalent comparison studies			
Curran A	21	17	0.87	1.02	−0.13	−0.05
Curran B	24	17	0.87	1.02	−0.13	−0.05
Curran C	20	17	0.97	1.02	−0.24	−0.05
Curran D	20	17	0.94	1.02	−0.27	−0.05
Dear	60	526	0.37	0.16	0.26	0.28
Dyer	225	193	0.36	0.19	0.44	0.38
French B	110	158	.	.	0.37	0.31
French C	161	158	.	.	0.51	0.31
FTC A	192	684	0.45	0.14	0.45	0.11
Keefauver	16	25	−0.03	0.17	0.49	0.30
Lass	38	82	0.64	0.53	0.44	0.41
Reynolds & Oberman	93	47	0.68	0.08	0.20	0.24
Zuman A	21	34	0.39	−0.17	0.52	−0.07
			Randomized control studies			
Alderman & Powers A	28	22	.	.	0.12	−0.10
Alderman & Powers B	39	40	.	.	0.04	−0.05
Alderman & Powers C	22	17	.	.	0.41	0.27
Alderman & Powers D	48	43	.	.	0.04	−0.10
Alderman & Powers E	25	74	.	.	0.13	0.14
Alderman & Powers F	37	35	.	.	0.31	0.17
Alderman & Powers G	24	70	.	.	0.60	0.42
Alderman & Powers H	16	19	.	.	0.04	0.03
Evans & Pike A	145	129	0.36	0.24	0.22	0.09
Evans & Pike B	72	129	0.49	0.24	0.15	0.09
Evans & Pike C	71	129	0.55	0.24	0.18	0.09
Laschewer	13	14	0.16	0.08	−0.06	−0.06
Roberts & Oppenheim A	154	111	.	.	0.08	−0.09
Roberts & Oppenheim B	188	122	0.02	−0.10	.	.
Zuman B	16	17	0.69	0.18	0.55	0.41
			Matched studies			
Burke A	25	25	.	.	0.86	0.36
Burke B	25	25	.	.	0.87	0.13
Coffin A	8	8	0.57	0.59	0.24	−0.60
Davis	22	21	0.00	−0.14	0.14	0.00
Frankel	45	45	1.16	0.81	0.66	0.53
Kintisch	38	38	.	.	0.41	0.35
Whitla	52	52	0.32	0.43	0.54	0.45

APPENDIX B
TABLE A-2
Results of coaching studies

Sample	Yr	Hrs	ETS	VI	MI	Vol	Sel	Itm	Tst	Tsk	Ex	Pb	Hwk	Ctr	Alp	Com
							Uncontrolled studies									
Coffin B	87	25	0	1	1	1	0	1	2	1	3	0	2	3	0	1
Coffman & Parry A	67	48	1	0	0	1	1	2	2	1	2	1	3	3	1	0
Coffman & Parry B	67	48	1	0	0	1	1	2	2	1	2	1	3	3	1	0
Johnson	84	30	0	1	1	2	0	1	2	3	3	0	3	3	2	0
Marron 1	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Marron 2	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Marron 3	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Marron 4	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Marron 5	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Marron 6	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Marron 7	65	100	1	1	1	2	2	3	3	3	2	0	1	3	2	0
Pallone A	61	45	0	1	0	2	2	1	2	1	2	1	3	3	1	0
Pallone B	61	100	0	1	0	1	2	1	2	1	2	1	3	3	1	0
							Nonequivalent comparison studies									
Curran A	88	6	0	1	1	1	2	1	2	2	2	0	2	3	0	1
Curran B	88	6	0	1	1	1	2	1	2	2	2	0	2	3	0	1
Curran C	88	6	0	1	1	1	2	1	2	2	2	0	2	3	0	1
Curran D	88	6	0	1	1	1	2	1	2	2	2	0	2	3	0	1
Dear	58	10	1	1	1	1	1	1	1	2	2	0	1	3	0	0
Dyer	53	15	1	1	1	0	2	1	1	2	2	1	1	3	0	0
French B	55	4.5	1	1	0	1	1	1	2	1	1	0	1	3	2	0
French C	55	15	1	1	1	1	1	1	1	2	2	0	1	3	2	0
FTC A[a]	78	40	0	1	1	2	1	1	1	1	2	0	1	3	1	0
Keefauver	76	14	0	1	1	0	2	1	1	1	1	0	1	3	1	0
Lass	61	·	1	1	1	1	1	3	3	3	3	0	3	3	0	0
Reynolds & Oberman	87	63	0	1	1	2	2	1	1	1	1	0	1	2	2	0
Zuman A	88	27	0	1	1	2	2	1	1	1	1	0	1	1	1	0

Sample	Yr	Hrs	ETS	VI	MI	Vol	Sel	Itm	Tst	Tsk	Ex	Pb	Hwk	Ctr	Alp	Com
							Randomized control studies									
Alderman & Powers A	80	7	1	1	0	2	1	1	1	3	3	1	3	1	1	0
Alderman & Powers B	80	10	1	1	0	2	1	1	2	3	3	1	3	2	1	0
Alderman & Powers C	80	10.5	1	1	0	2	1	1	2	3	3	1	3	1	1	0
Alderman & Powers D	80	10	1	1	0	2	1	1	2	3	3	1	3	2	1	0
Alderman & Powers E	80	6	1	1	0	2	1	1	2	3	3	1	3	1	1	0
Alderman & Powers F	80	5	1	1	0	0	2	1	2	3	3	1	3	1	1	0
Alderman & Powers G	80	11	1	1	0	2	2	1	2	3	3	1	3	1	1	0
Alderman & Powers H	80	45	1	1	0	1	2	1	2	3	3	1	3	1	1	0
Evans & Pike A	73	21	1	0	1	2	1	1	2	1	2	1	1	1	1	0
Evans & Pike B	73	21	1	0	1	2	1	1	2	1	2	1	1	1	1	0
Evans & Pike C	73	21	1	0	1	2	1	1	2	1	2	1	1	1	1	0
Laschewer	86	8.9	0	1	1	2	2	1	2	2	2	0	2	1	0	1
Roberts & Oppenheim A	66	7.5	1	1	0	2	0	1	2	1	2	0	2	1	1	0
Roberts & Oppenheim B	66	7.5	1	0	1	2	0	1	2	1	2	0	2	1	1	0
Zuman B	88	24	0	1	1	1	1	1	1	1	1	0	1	1	1	0
							Matched studies									
Burke A	86	50	0	1	0	1	1	1	1	1	1	0	1	3	2	0
Burke B	86	50	0	1	0	1	1	1	1	1	1	0	1	3	2	0
Coffin A	87	18	0	1	1	1	0	1	2	2	1	0	2	2	0	1
Davis	85	15	0	1	1	1	2	1	2	2	2	0	2	1	1	1
Frankel	60	30	0	1	1	2	2	1	3	3	3	1	3	3	1	0
Kintisch	79	.	0	1	0	1	2	1	2	1	1	1	1	3	1	0
Whitla	62	10	1	1	1	2	0	1	1	1	2	1	1	3	2	0

Note. Predictor variables are publication date (Yr), duration (Hrs), ETS sponsorship (ETS), presence of verbal instruction (VI), presence of math instruction (MI), voluntariness (Vol), selectivity (Sel), presence of item practice (Itm), presence of test practice (Tst), presence of test-taking skill instruction (Tsk), use of extra activities (Ex), whether the study was published (Pb), presence of homework (Hwk), the type of control group (Ctr), instruction for alpha abilities (Alp), and use of computerized coaching (Com).

[a] Descriptive data for the FTC study were taken from Rock (1980).

APPENDIX C

This appendix provides formulas for the variances and covariances of the $\hat{\Delta}$ values (the differences in standardized mean-change measures).

Variances

Becker (1988) presented asymptotic distributions for the standardized mean change and for differences (i.e., $\hat{\Delta}$ values) between standardized mean changes for experimental and control samples. For a study of coaching that has n^C subjects in the coached group and n^U in the control group, approximately unbiased estimates of the two standardized mean changes are

$$d^C = \frac{4(n^C - 2)}{4n^C - 5}\left(\frac{\bar{Y}^C - \bar{X}^C}{S_y^C}\right)$$

and

$$d^U = \frac{4(n^U - 2)}{4n^U - 5}\left(\frac{\bar{Y}^U - \bar{X}^U}{S_y^U}\right),$$

where $\bar{X}^C$ and $\bar{Y}^C$ are pretest and posttest SAT means and S_y^C is the posttest standard deviation for the coached sample and $\bar{X}^U$, $\bar{Y}^U$, and S_y^U are analogous statistics for the uncoached sample.

The difference in standardized mean changes is $\hat{\Delta} = d^C - d^U$, and the estimated variance of $\hat{\Delta}$ is

$$V(\hat{\Delta}) = \frac{4(1 - r^C) + (d^C)^2}{2n^C} + \frac{4(1 - r^U) + (d^U)^2}{2n^U},$$

$$= V(d^C) + V(d^U),$$

where r^C and r^U are estimators of the pretest-posttest correlations (i.e., values of r_{xy}, which in this synthesis are correlations between pre- and posttest SAT scores) for the coached and uncoached samples, respectively.

The value of $r = 0.88$ was used for the pretest-posttest correlation for both SAT-M and SAT-V for all subjects, following DerSimonian and Laird (1983).

When $\tilde{\Delta}$ values are computed for studies without control groups, their variances depend on the variance of $d.^U$, the "imputed" control average change. The variance of $\tilde{\Delta}$ is

$$V(\tilde{\Delta}) = V(d^C) + V(d.^U),$$

where $d.^U$ is the weighted average of the standardized mean-change measures for all existing control groups and $V(d.^U)$ is the estimated variance of that mean. If there are k existing control results (on the outcome of interest) and $V(d_i^U)$ is the variance of the mean change in the ith control group, then

$$V(d.^U) = \left[\sum_{i=1}^{k} \frac{1}{V(d_i^U)}\right]^{-1}$$

Covariances

Table 3 in the text lists the 10 types of interrelationships existing among the coaching studies. Although the particular formulas for the 10 covariances differ somewhat, they are all obtained using the basic rules of the algebra of expectations.

Details are shown for the first covariance, but only the final formulas are given for the following nine. These are presented in Table A-3.

Covariance between SAT-M results of two uncontrolled studies. This covariance is also equal to the covariance between SAT-M results of one uncontrolled and one comparison study. Below are the details of the derivation for two uncontrolled studies. Let $\tilde{\Delta}_i^{M}$ and $\tilde{\Delta}_j^{M}$ be the "imputed" $\hat{\Delta}$ values on SAT-M for the ith and jth studies. The covariance $\mathrm{Cov}(\tilde{\Delta}_i^{M}, \tilde{\Delta}_j^{M})$ is of interest.
Then

$$\mathrm{Cov}(\tilde{\Delta}_i^{M}, \tilde{\Delta}_j^{M})$$

$$= \mathrm{Cov}([d_i^{CM} - d_{\cdot}^{UM}], [d_j^{CM} - d_{\cdot}^{UM}])$$

$$= \mathrm{Cov}(d_i^{CM}, d_j^{CM}) - \mathrm{Cov}(d_i^{CM}, d_{\cdot}^{UM}) - \mathrm{Cov}(d_{\cdot}^{UM}, d_j^{CM}) + \mathrm{Cov}(d_{\cdot}^{UM}, d_{\cdot}^{UM}).$$

Of these four terms, the first three are equal to zero because the standardized mean changes d_i^{CM} and d_j^{CM} are from independent uncontrolled studies (i.e., no d_i^{UM} values were included in computing $d_{\cdot}^{UM}$). Thus,

$$\mathrm{Cov}(\tilde{\Delta}_i^{U}, \tilde{\Delta}_j^{M}) = \mathrm{Cov}(d_{\cdot}^{UM}, d_{\cdot}^{UM}) = V(d_{\cdot}^{UM}).$$

TABLE A-3
Covariances among study results

Label[a]	Covariance
A	$\mathrm{Cov}(\tilde{\Delta}_i^{M}, \tilde{\Delta}_j^{M}) = V(d_{\cdot}^{UM})$
B	$\mathrm{Cov}(\tilde{\Delta}_i^{V}, \tilde{\Delta}_j^{V}) = V(d_{\cdot}^{UV})$
C	$\mathrm{Cov}(\tilde{\Delta}_i^{M}, \tilde{\Delta}_j^{V}) = r_{MV}\, V(d_{\cdot}^{UM})\, V(d_{\cdot}^{UV}) \sum_{s=1}^{k} \frac{1}{\sqrt{V(d_s^{UM})\, V(d_s^{UV})}}$
D	$\mathrm{Cov}(\tilde{\Delta}_i^{M}, \hat{\Delta}_j^{V}) = r_{MV}\, V(d_{\cdot}^{UM}) \sqrt{V(d_j^{UV})/V(d_j^{UM})}$
E	$\mathrm{Cov}(\tilde{\Delta}_i^{V}, \hat{\Delta}_j^{V}) = r_{MV}\, V(d_{\cdot}^{UV}) \sqrt{V(d_j^{UM})/V(d_j^{UV})}$
F	$\mathrm{Cov}(\hat{\Delta}_i^{M}, \hat{\Delta}_j^{M}) = \mathrm{Cov}([d_i^{CM} - d_s^{UM}], [d_j^{CM} - d_s^{UM}]) = V(d_s^{UM})$
G	$\mathrm{Cov}(\hat{\Delta}_i^{V}, \hat{\Delta}_j^{V}) = \mathrm{Cov}([d_i^{CV} - d_s^{UV}], [d_j^{CV} - d_s^{UV}]) = V(d_s^{UV})$
H	$\mathrm{Cov}(\hat{\Delta}_i^{M}, \hat{\Delta}_j^{V}) = \mathrm{Cov}([d_i^{CM} - d_s^{UM}], [d_j^{CV} - d_s^{UV}]) = \mathrm{Cov}(d_s^{UM}, d_s^{UV})$
I	$\mathrm{Cov}(\tilde{\Delta}_i^{M}, \tilde{\Delta}_i^{V}) = r_{MV}\sqrt{V(d_i^{CM})\, V(d_i^{CV})} + r_{MV}\sqrt{V(\mathrm{d}_{\cdot}^{UM})\, V(d_{\cdot}^{UV})} \sum_{s=1}^{k} \frac{1}{\sqrt{V(d_s^{UM})\, V(d_s^{UV})}}$
J	$\mathrm{Cov}(\hat{\Delta}_i^{M}, \hat{\Delta}_i^{V}) = r_{MV}\,[\sqrt{V(d_i^{CM})\, V(d_i^{CV})} + \sqrt{V(d_i^{UM})\, V(d_i^{UV})}]$ and $\mathrm{Cov}(\hat{\Delta}_i^{M}, \hat{\Delta}_i^{V}) = \mathrm{Cov}([d_i^{CM} - d_s^{UM}], [d_i^{CV} - d_s^{UV}]) = r_{MV}\,[\sqrt{V(d_i^{CM})\, V(d_i^{CV})} + \sqrt{V(d_s^{UM})\, V(d_s^{UV})}]$

Note. r_{MV} is the correlation between SAT-M and SAT-V scores.
[a] See Table 3 for label descriptions.

References

* Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-verbal scores. *American Educational Research Journal, 17*, 239–253.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257–278.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*, 1262–1264.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 429–444). New York: American Council on Education and Macmillan Publishing Co.

* Burke, K. B. (1986). *A model reading course and its effects on the verbal scores of eleventh and twelfth grade students on the Nelson Denny Test, the Preliminary Scholastic Aptitude Test, and the Scholastic Aptitude Test.* Unpublished doctoral dissertation, Georgia State University. (University Microfilms No. 86-26152).

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally.

* Coffin, G. C. (1987, February). *Computer as a tool in SAT preparation.* Paper presented at the annual meeting of the Florida Instructional Computing Conference, Orlando, FL. (ERIC Document Reproduction Service No. ED 286 932).

* Coffman, W. E., & Parry, M. E. (1967). Effects of an accelerated reading course on SAT-V scores. *Personnel and Guidance Journal, 46*, 292–296.

Cole, N. (1982). The implications of coaching for ability testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences and controversies* (pp. 389–414). Washington, DC: National Academy Press.

College Entrance Examination Board. (1983). *10 SATs: Scholastic Aptitude Tests of the College Board.* New York: Author.

* Curran, R. G. (1988). *The effectiveness of computerized coaching for the Preliminary Scholastic Aptitude Test (PSAT/NMSQT) and the Scholastic Aptitude Test (SAT).* Unpublished doctoral dissertation, Boston University. (University Microfilms No. 88-14377)

* Davis, W. D. (1985). *An empirical assessment of selected computer software purported to raise SAT scores significantly when utilized with short-term computer-assisted instruction on the microcomputer.* Unpublished doctoral dissertation, Florida State University. (ERIC Document Reproduction Service No. ED 283 370)

* Dear, R. E. (1958). *The effect of a program of intensive coaching on SAT scores* (ETS RB 58-5). Princeton, NJ: Educational Testing Service.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*, 1–15.

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

* Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10*, 257–272.

Federal Trade Commission, Boston Regional Office. (1978). *Staff memorandum of the Boston Regional Office of the Federal Trade Commission: The effects of coaching on standardized admission examinations* (NTIS No. PB-296 210). Boston: Author.

Federal Trade Commission, Bureau of Consumer Protection. (1979). *Effects of coaching on standardized admission examinations: Revised statistical analyses of data gathered by Boston Regional Office of the Federal Trade Commission* (NTIS No. PB-296 196). Washington, DC: Author.

* Frankel, E. (1960). Effects of growth, practice, and coaching on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 38*, 713–719.

* French, J. W. (1955). *The coachability of the SAT in public schools* (ETS RB 55-26). Princeton, NJ: Educational Testing Service.

French, J. W., & Dear, R. E. (1959). Effect of coaching on an aptitude test. *Educational and Psychological Measurement, 19*, 319–330.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*, 3–8.

Hedges, L. V. (1986). Issues in research synthesis. *Review of Research in Education, 13*, 353–398.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Jackson, R. (1980). The Scholastic Aptitude Test: A response to Slack and Porter's "Critical Appraisal." *Harvard Educational Review, 50*, 382–391.

* Johnson, S. T. (1984). *Preparing Black students for the SAT—Does it make a difference?* (An evaluation report of the NAACP Test Preparation Project). Unpublished report to the National Association for the Advancement of Colored People, New York. (ERIC Document Reproduction Service No. ED 247 350)

* Keefauver, L. W. (1976). *The effects of a program of coaching on Scholastic Aptitude Test scores of high school seniors pretested as juniors.* Unpublished doctoral dissertation, University of Tennessee at Knoxville. (University Microfilms No. 77-3651).

* Kintisch, L. S. (1979). Classroom techniques for improving Scholastic Aptitude Test scores. *Journal of Reading, 22*, 416–419.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179–188.

* Laschewer, A. D. (1986). *The effect of computer assisted instruction as a coaching technique for the Scholastic Aptitude Test preparation of high school juniors.* Unpublished doctoral dissertation, Hofstra University. (University Microfilms No. 86-06936)

* Lass, A. H. (1961). Unpublished report. (Cited in Pike, L. W. (1978). *Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with recommendations.* New York: College Entrance Examination Board.)

Lewis, R. (1989, November 23). Questions still persist over SATs. *The Flint Journal*, p. 67.

* Marron, J. E. (1965). *Preparatory school test preparation: Special test preparation, its effect on College Board scores and the relationship of affected scores to subsequent college performance.* West Point: Research Division, Office of the Director of Admissions and Registrar, United States Military Academy.

Mehrens, W. A., & Kaminski, J. (1988, April). *Using commercial test preparation materials for improving standardized test scores: Fruitful, fruitless or fraudulent?* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Messick, S. (1980). *The effectiveness of coaching for the SAT: Review and analysis of research from the fifties to the FTC* (Research report No. 80-8). Princeton, NJ: Educational Testing Service.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216.

National School Boards Association. (1984). Raise expectations to achieve excellence. *Updating School Board Policies, 15*, 1–8.

Orwin, R. G., & Cordray, D. S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Journal of Applied Psychology, 97*, 134–147.

Owen, D. B. (1985). *None of the above: Behind the myth of scholastic aptitude.* Boston: Houghton Mifflin.

Owens, P. (1983). Computer coaching of the SAT. *Popular Computing, 2*, 186–188, 190, 192, 194, 198.

* Pallone, N. J. (1961). Effects of short-term and long-term developmental reading courses upon SAT verbal scores. *Personnel and Guidance Journal, 39*, 654–657.

Pike, L. W. (1978). *Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with recommendations.* New York: College Entrance Examination Board.

Powell, B., & Steelman, L. C. (1984). Variations in state SAT performance: Meaningful or

misleading? *Harvard Educational Review, 54*, 389–412.

Powers, D. E. (1982). *Estimating the effects of various methods of preparing for the SAT* (Research Report No. 82-23). New York: College Entrance Examination Board.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1987). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111–120.

* Reynolds, A. J., & Oberman, G. O. (1987, April). *An analysis of a PSAT preparation program for urban gifted students.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

* Roberts, S. O., & Oppenheim, D. B. (1966). *The effect of special instruction upon test performance of high school students in Tennessee* (CB RDR 66-7, No. 1, and ETS RB 66-36). Princeton, NJ: Educational Testing Service.

* Rock, D. A. (1980). Disentangling coaching effects and differential growth in the FTC coaching study. In S. Messick (Ed.), *The effectiveness of coaching for the SAT: Review and analysis of research from the fifties to the FTC* (Research Report No. 80-8, pp. 123–135). Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Slack, W. V., & Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review, 50*, 154–175.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation in Education: An International Review Series, 4*, 22–24.

Staples, B. (1985, April). SAT packages—An update. *Creative Computing*, pp. 86–89.

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

Winokur, H. (1983). *The effects of special preparation for the verbal section of the SAT.* Unpublished doctoral dissertation, Virginia Polytechnic and State University.

* Zuman, J. P. (1988, April). *The effectiveness of special preparation for the SAT: An evaluation of a commercial coaching school.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 294 900)

---

* Data from these references were included in the statistical analyses of SAT coaching study results.

Author

BETSY JANE BECKER is Associate Professor, College of Education, Michigan State University, 456 Erickson Hall, East Lansing, MI 48824-1034. She specializes in research synthesis, quantitative methods, and gender differences in mathematics and science.