META-ANALYSIS

A Case Study

DEREK C. BRIGGS
University of Colorado, Boulder

This article raises some questions about the usefulness of meta-analysis as a means of reviewing quantitative research in the social sciences. When a meta-analytic model for SAT coaching is used to predict results from future studies, the amount of prediction error is quite large. Interpretations of meta-analytic regressions and quantifications of program and study characteristics are shown to be equivocal. The match between the assumptions of the meta-analytic model and the data from SAT coaching studies is not good, making statistical inferences problematic. Researcher subjectivity is no less problematic in the context of a meta-analysis than in a narrative review.

Keywords: *meta-analysis; literature review; SAT coaching; statistical inference*

Literature reviews play a critical role in research. Through the analysis and evaluation of past research on a particular topic, the stage is set for new theoretical and empirical contributions. Since the publication of *Meta-Analysis in Social Research* (Glass, McGaw, and Smith 1981) and *Statistical Methods for Meta-Analysis* (Hedges and Olkin 1985) in the early 1980s, the meta-analysis has become an accepted and, in many instances, a preferred methodological approach for the review of quantitative research studies in educational research. A search for the keyword *meta-analysis* in the Educational Resources Information Center (ERIC) between 1980 and 2003 turns up well more than 1,000 citations of articles from peer-reviewed research journals or conference presentations. Meta-analysis has been used to review topics such as the effects of rewards on intrinsic motivation, the relationship between educational resources and achievement, the effectiveness of phonics instruction, and the effectiveness of bilingual education. In the textbook *Experimental and Quasi-Experimental Design for Generalized Causal Inference*, the authors describe meta-analysis as "one of the most important social

AUTHOR'S NOTE: *The author thanks David Freedman and Lorrie Shepard for helpful comments on earlier versions of this article.*

EVALUATION REVIEW, Vol. 29 No. 2, April 2005 87-127
DOI: 10.1177/0193841X04272555

science developments in the second half of the 20th century" (Shadish, Cook, and Campbell 2002, 446).

The use of meta-analysis has been criticized on statistical grounds in the past (Oakes 1986; Wachter 1988; Berk 2004; Berk and Freedman 2003). Indeed, the potential for the misuse of meta-analysis has been well recognized even by those who have done the most to develop it as a methodology (Hedges 1990; Kulik and Kulik 1988). A key issue, and one that I revisit here, is whether the fundamental assumptions of the meta-analytic model are likely to hold for the data under review. The primary purpose in this article is to raise some empirically grounded questions about the usefulness of the meta-analytic approach. I will suggest that a meta-analysis is ultimately not that much different from a systematic narrative review, but with the unfortunate distinction that the role of human judgment in the meta-analysis is more easily obscured.

This article is not a meta-meta-analysis but a case study of a specific application of meta-analysis in a review of educational research. In this article, I critique a meta-analysis used to synthesize evidence about the effectiveness of coaching programs for the SAT. Meta-analyses are inherently difficult to critique for two reasons: (a) the prohibitive cost of reading and reviewing all the studies that form the basis of the meta-analysis and (b) a lack of new studies available to test the validity of the meta-analytic model. Because I have gathered, read and evaluated virtually[1] every published and unpublished study of SAT coaching that has been the subject of previous reviews as background for a different project (Briggs 2001, 2002b, 2004a, 2004b), and because many new studies have been conducted since the last major review of the SAT coaching literature, I find myself in a good position to evaluate the merits of the meta-analytic approach in this restricted context. There is no statistical basis for generalizing the context-specific findings here to all meta-analyses, but I believe that the sort of data examined here are typical of the sort found in other social science research contexts. And although the context and methods for a meta-analysis may vary, the basic steps of the approach tend to be the same. In all likelihood, the issues raised in this article are ones that any meta-analysis would need to address.

There are four principal sections to this article. In the first section, I provide the necessary context for my case study. I give some background on SAT coaching studies and provide a frame of reference for the interpretation of estimated coaching effects. I next introduce a widely cited and well-regarded meta-analysis of SAT coaching studies conducted by Betsy Jane Becker in 1990. In the second section, I evaluate Becker's meta-analytic regression models by using them to predict the results of 24 new studies that were not included in the 1990 review. In the third section, I analyze the way that studies have been quantified in Becker's meta-analysis. I present the fundamental assumptions behind Becker's meta-analytic model and consider the plausibility of these assumptions relative to the data found in the underlying studies. In the fourth section, I compare the conclusions reached by Becker's meta-analysis to those one might reach after conducting a systematic narrative review.

SAT COACHING AND META ANALYSIS

BACKGROUND ON SAT COACHING

Coaching can be defined as systematic test preparation for a student or group of students that involves the following: content review, item drill and practice, and an emphasis on specific test-taking strategies and general test wiseness. (For more detailed discussions, see Pike 1978; Anastasi 1981; Cole 1982; Messick 1982; Bond 1989). Coaching is commonly offered for the SAT, one of the most widely known college admissions tests in the United States. The SAT has traditionally consisted of two sections administered over two and a half hours, with items that measure, respectively, verbal (SAT-V) and mathematical (SAT-M) reasoning ability.[2] The SAT is designed to measures reasoning abilities that are developed gradually over the years of primary and secondary schooling that precede college (Messick 1980; Anastasi 1981). Companies such as Kaplan and The Princeton Review claim they can improve a student's SAT performance through coaching programs that are short term in nature.

The SAT is reported on a scale ranging from 200 to 800 points per test section. Because the SAT scale has no absolute meaning, the best way to interpret the effect size of a coaching treatment is relative to the standard deviation *(SD)* of each test section, which is about 100 points. Hence, a 10-point coaching effect on one section of the SAT would be equivalent to an effect size of 0.1 of an *SD*, a relatively small effect. A coaching effect of 60 points, on the other hand, is equivalent to an effect size of 0.6 of an *SD*, a relatively large effect.

Between 1953 and 1993, there were about 30 reports on the effectiveness of SAT coaching. Twelve reviews of coaching reports were written between 1978 and 1993. Nine of them are narrative reviews (Pike 1978; Messick 1980; Slack and Porter 1980; Anastasi 1981; Messick and Jungeblut 1981; Cole 1982; Messick 1982; Bond 1989; Powers 1993). The other 3 reviews are meta-analyses (DerSimonian and Laird 1983; Kulik, Bangert-Drowns, and

Kulik 1984; Becker 1990). Becker's (1990) review stands out from the crowd. This review involved the synthesis of 25 reports on SAT coaching drawn from academic journals, doctoral dissertations, institutional reports, and conference papers. (Key characteristics of the reports that form the basis for Becker's review, as well as 16 others that were not, are summarized in Appendix A.) Becker's encompassing retrieval of coaching reports was paired with state-of-the-art meta-analytic techniques. Since its publication, Becker's review has been cited in more than 20 other published articles on related topics.

AN OVERVIEW OF BECKER'S META-ANALYSIS OF SAT COACHING STUDIES

Meta-analysis starts with a collection of summary data from a series of studies, each of which is considered a replication of the same basic experiment. As defined here, a *study* constitutes any single estimate of the effect of coaching on either the SAT-M or SAT-V. Multiple studies (usually at least two) are typically found within what I will term a *report* on SAT coaching. I will index reports with the subscript *h*. Within each report *h*, I find one or more studies, indexed by the subscript *i*. In Becker's 1990 review, there are a total of 25 reports that contain 70 studies. Students are indexed by the subscript *j*. The treatment of interest is whether student *j* gets coached after taking the SAT a first time. Let X_{hij}^{C} and Y_{hij}^{C} represent the SAT scores for coached student *j* in study *i* within report *h*. This student takes the SAT twice and receives coaching in between testings ("C"). On the other hand, let X_{hij}^{U} and Y_{hij}^{U} represent the SAT scores for uncoached student *j* in study *i* within report *h*. This student takes the SAT twice but never receives coaching ("U").

The outcome variable in the meta-analysis is the effect size of the coaching treatment. To calculate this, Becker first computes the standardized mean change in SAT scores for coached and uncoached students. For each group, the standardized mean change is simply the difference between mean posttest and pretest SAT scores divided by the posttest standard deviation:

$$g_{hi}^{C} = \frac{\left(\overline{Y}_{hi}^{C} - \overline{X}_{hi}^{C}\right)}{SD^{C}Y_{hi}}. \tag{1}$$

$$g_{hi}^{U} = \frac{\left(\overline{Y}_{hi}^{U} - \overline{X}_{hi}^{U}\right)}{SD^{U}Y_{hi}}. \tag{2}$$

From this, Becker computes the effect size estimate[3]

$$\hat{\Delta}_{hi} = g_{hi}^{C} - g_{hi}^{U} . \qquad (3)$$

Using this approach, one way to get an overall estimate of the coaching effect would be to calculate a weighted average of $\hat{\Delta}_{hi}$. I refer to this as Model A. In Becker's meta-analysis, three other regression models are to adjust effect size estimates for differences across studies. I refer to these as Models B, C, and D. Below is a summary of the central features of these models. (The specific variables used in Models C and D are presented in more detail later.)

Model A: No predictor variables
Model B: One predictor variable indicating whether the study was based on math or verbal SAT scores
Model C: Set of predictor variables intended to characterize the nature of coaching in the study
Model D: Set of predictor variables intended to characterize the design of the study

TESTING BECKER'S SAT COACHING META-ANALYSIS WITH NEW STUDIES

A good way to test the usefulness of a model is to attempt to validate it with new data. Another 16 coaching reports have come to light since Becker's meta-analysis was published.[4] I found these reports by searching through the Web, academic journals, and the University Microfilms index of dissertation abstracts using combinations of the keywords *SAT*, *coaching*, and *test preparation*. For narrative summaries of these reports, see Briggs (2002a). From each of these new reports, one can get one or more estimates of a coaching effect from studies within the report. Each of these studies is listed in the rows of Table 1. I coded each of these studies using the same variables Becker created for her meta-analysis (see Appendixes B and C for details). The last four columns of Table 1 give the coaching effect predicted for each study as a function of its coaching content and design characteristics, using the specifications represented by Models A through D.

Table 2 summarizes the prediction error for Models A through D using the root mean square error (RMSE). The lowest RMSE for studies estimating coaching effects for either section of the SAT is about 14 points (Model A), the highest is about 29 (Model D). These prediction errors are about the same

TABLE 1: Observed and Predicted Effects From New Coaching Studies

			Predicted Coaching Effect From Becker (1990)			
Report	*Study*	*Coaching Effect*	*Model A*	*Model B*	*Model C*	*Model D*
Hopmeier (1984)	SAT-V	57	30	11.6	–12.9	24.5
	SAT-M	37	30	25.5	–1.2	35.8
Fraker (1987)	SAT-V	–16	30	11.6	1.9	0.8
	SAT-M	16	30	25.5	13.6	12.1
Harvey (1988)	SAT-M	21	30	25.5	–14.5	8.1
Whitla (1988)	SAT-V	11	30	11.6	2.7	0.5
	SAT-M	16	30	25.5	14.4	11.8
Snedecor (1989)	SAT-V	0	30	11.6	2.7	–0.2
	SAT-M	15	30	25.5	14.4	11.1
Wing, Childs, and Maxwell (1989)	SAT-V	4	30	11.6	–0.7	–0.6
	SAT-M	33	30	25.5	11	10.7
Smyth (1990)	SAT-V	9	30	11.6	2.7	–0.9
	SAT-M	18	30	25.5	14.4	10.4
Shaw (1992)	SAT-V	21	30	11.6	–27.3	17.8
	SAT-M	6	30	25.5	–15.6	29.1
Schroeder (1992)	SAT-M	46	30	25.5	–3.8	–11
Holmes and Keffer (1995)	SAT-V	39	30	11.6	–26.2	15.2
Wrinkle (1996)	SAT-V	31	30	11.6	44.7	27.8
Powers and Rock (1999)	SAT-V	6	30	11.6	2.7	–23.7
	SAT-M	18	30	25.5	14.4	–12.4
Briggs (2001)	SAT-V	11	30	11.6	2.7	–5.3
	SAT-M	19	30	25.5	14.4	6
Kaplan (2002)	SAT-M (Year 1)	37	30	25.5	8.6	–7.5
	SAT-M (Year 2)	59	30	25.5	12.1	–4.5

NOTE: Model A = Column 1 in Becker's Table 7; Model B = Column 2; Model C = Column 3; Model D = Column 4.

TABLE 2: Average Prediction Error From Becker's (1990) Meta-Analytic Models

	Root Mean Square Error for Predicted Coaching Effect	
Meta-Analytic Model	*SAT-V*	*SAT-M*
A: No Predictors	24.0	13.6
B: Math/Verbal Difference	21.6	17.3
C: Coaching Content	27.7	28.5
D: Study Design	17.4	28.6

magnitude as most of the coaching effects estimated in individual studies. It seems that Becker's meta-analysis does poorly when it comes to predicting the results of new studies.[5] To make this more concrete, when Model D is used to predict the results of new coaching studies, one must expect predictions of SAT-V coaching effects to be off on average by +/–17 points. Predictions of SAT-M effects, using the same model, will be off on average by +/–29 points. Note that the RMSE does not decrease once treatment and design characteristics of the studies are controlled. This is cause for concern, because the reason these sorts of regression adjustments are made is to better predict the effect of coaching. The next section shows why such adjustments fall short.

INTERPRETING META-ANALYTIC REGRESSIONS AND QUANTIFYING STUDIES

INTERPRETING META-ANALYTIC REGRESSIONS

Table 3 presents the results from Models A through D, estimated using generalized least squares. These results correspond to Table 7 of Becker's 1990 review (p. 393). For ease of interpretation, I have translated the estimated regression coefficients from effect size units back into SAT scores. (I assume that the relevant *SD* for each section of the SAT is 100 points.)

Models C and D include predictor variables meant to control for differences in coaching studies. In Model C, the emphasis is on controlling for differences in coaching programs. The model includes predictor variables that indicate whether coaching emphasized verbal instruction, math instruction, alpha instruction,[6] item practice, test practice, test-taking skills, other preparatory activities, homework, and computer instruction. Model C also

TABLE 3: Becker's (1990) Meta-Analytic Regressions for 70 Coaching Study Outcomes

	Model A	*Model B*	*Model C*	*Model D*
Predictor	*No Predictors*	*Math/Verbal Difference*	*Coaching Content*	*Study Design*
Grand mean	30.0*	25.5*	–38.8*	40.9*
SAT-M		11.6*	11.7	11.3*
Control group			–2.6	–1.0
Duration			0.7*	0.6*
Verbal instruction			16.3*	20.2*
Math instruction			–4.1	
Alpha instruction			0.8	
Item practice			23.5*	
Test practice			–2.0	
Test-taking skills			–0.9	
Other activities			–3.4	
Homework			0.0	
Computer instruction			–7.5	
Wait-list control			8.3	
Alternative control			–6.9	
Year				–0.7*
Publication type				0.4
Use of matching				1.5
Use of randomization				23.8*
ETS sponsorship				–19.8
Selectivity				–3.3
Voluntariness				–0.4

SOURCE: Becker (1990), Table 7.
NOTE: Effect sizes have been placed on the SAT scale assuming population *SD*s of 100 points.
*Slope coefficients for predictors are statistically significant at $\alpha = .05$.

includes dummy variables that indicate whether a study involved a control group of uncoached students ("Control Group"), whether the control group derived from a wait-list of students interested in receiving the coaching treatment ("Wait-List Control"), and whether the control group received an alternate form of coaching ("Alternate Control").[7] The only continuous variable measures the duration of the coaching treatment in hours.

In Model D, the emphasis is on controlling for differences in study designs. However, there is overlap between the variables specified here and the variables specified in Model C. Among the new variables are predictors indicating the year the SAT coaching report was completed and whether the report had been published, made use of random assignment to treatment conditions, employed statistical matching, and was sponsored by Educational Testing Service (ETS). Ordinal variables ranging in values from 0 to 2 were included for sample selectivity (“Selectivity”) and motivation (“Voluntariness”). For Selectivity, 0 represents a study sample comprised of students with low academic ability, 1 represents a sample of students with mixed academic ability, and 2 represents a sample of students with high academic ability. For Voluntariness, 0 represents compulsory participation in a coaching program, 1 represents participation in a coaching program possible with little cost, and 2 represents participation in a coaching program that is voluntary. Studies with higher values on the Voluntariness variable are those where coached students are better motivated.

In Model A, no predictor variables are included in the regression. The coefficient for the grand mean, 30, is the weighted average of coaching effects across all studies, for both SAT-V and SAT-M. The total effect of coaching on the SAT is predicted to be 60 points. In Model B, a dummy variable is included for SAT-M. The results suggest that the SAT-V coaching effect is about 26 points, whereas the SAT-M effect is about 37 points. Again, the total effect of coaching appears to be about 60 points.

According to Model C, coaching duration, verbal instruction, and item practice are significant predictors of effect size. Imagine that there are two coaching studies that only differ on these three variables. The first study involves a coaching program that is 10 hours long with no verbal instruction or item practice. By contrast, a second coaching study involves a coaching program 20 hours long with verbal instruction and item practice. Using Model C, the predicted SAT-V and SAT-M coaching effects for the second coaching study would be 46 and 47 points higher than those predicted for the first coaching study. Should one conclude that coaching of longer duration with verbal instruction and item practice is demonstrably more effective than coaching of shorter duration without verbal instruction and item practice? According to the regression model, the answer is yes. However, if one looks more closely at the underlying data, it turns out that almost *all* coaching programs under review involve item practice and verbal instruction. There are 70 studies represented in Table 3. Of these 70 studies, only 2 come from programs that did not involve item practice; only 6 come from programs that did not involve verbal instruction. Only one study (Coffman and Parry 1967) had neither verbal instruction nor item practice. It is this last study that is driving the results of the regression model. The treatment in the Coffman and Parry (1967) study was in fact a speed-reading course offered to university freshmen, making it a remarkably atypical SAT coaching study to drive these results.

Model D is meant to control for differences in study designs. According to Model D, all else held constant, the predicted coaching effect on either section of the SAT is 24 points higher for studies with randomized experimental designs than for studies with nonrandomized designs. There are five reports with randomized designs in Becker's review. The estimated coaching effects in these reports are listed in Table 4. By inspection, the estimated coaching effects are not especially large, with the exception of Zuman (1988). The median estimated SAT-M and SAT-V coaching effects for studies with randomized designs are only 12 and 14 points. There appear to be substantial differences between these observed coaching effects and the effects predicted by Becker's Model D.

Becker (1990) explains this discrepancy as follows:

> 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. (P. 396)

The sort of adjustment being referenced is accomplished in meta-analysis by including the sorts of "control" variables found in Model D. To see why adjustments might be necessary, one can compare the average characteristics for randomized and nonrandomized studies. These characteristics are summarized in Table 5.

As Table 5 indicates, the characteristics of randomized studies are considerably different from those of nonrandomized studies. Randomized studies are much more likely to be published and to be sponsored by ETS. Nonrandomized studies involve coaching that is more than twice as long in duration as in randomized studies. They are typically not sponsored by ETS. These differences in the characteristics of the two groups of studies may serve to confound any comparisons between the two, so it might seem sensible to hold these characteristics constant. Unfortunately, it is rather easy to lose sight of the empirical assumptions required when we hold confounding variables constant. Figure 1 brings this into sharper relief.

Figure 1 plots Becker's estimated coaching effect per study as a function of coaching duration. Randomized studies are represented by solid circles, nonrandomized studies by empty circles. For each group, separate lines for the regression of coaching effect on duration are shown. The dashed line represents the regression line for randomized studies; the solid line is for nonrandomized studies. The slopes of these two lines are clearly different. With randomized studies, there is little relationship between coaching effects and duration. With nonrandomized studies, there is a strong positive associa-

TABLE 4: Estimated Coaching Effects in Randomized Studies

Report and Study	*SAT-M*	*SAT-V*
Alderman and Powers (1980)		
School A		22
School B		9
School C		14
School D		14
School E		–1
School F		14
School G		18
School H		1
Evans and Pike (1973)		
Group A	12	
Group B	25	
Group C	11	
Laschewer (1985)	8	0
Roberts and Openheim (1966)		
School A		17
School B	12	
Zuman (1988)	51	14
Median effect estimate	12	14

TABLE 5: Design Characteristics for Randomized and Nonrandomized Coaching Studies

Design Characteristic	*Randomized Studies*	*Nonrandomized Studies*
Control group	100%	70%
Mean duration	15.6 hours	36.2 hours
Verbal instruction	65%	96%
Mean year study released	1978	1974
Published	70%	22%
Matching	0%	22%
Educational Testing Service sponsorship	80%	38%
Selectivity (0-2)	1.15	1.44
Voluntariness (0-2)	1.75	1.34
Number of studies	20	50

tion.[8] The grand mean for coaching duration among all studies is about 30 hours (represented in Figure 1 by a vertical dashed line). With the exception of one study, the duration of coaching for randomized studies is less than 30 hours. The regression adjustment of Model D is based on the assumption

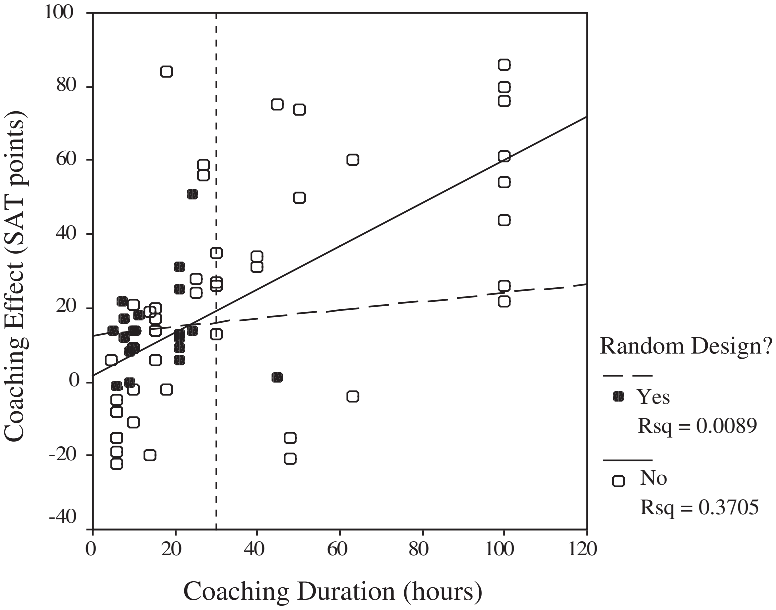

Figure 1: Coaching Effects by Duration for Randomized and Nonrandomized Designs

that, once other design characteristics are held constant, the two regression lines in Figure 1 will be parallel, with the line for the randomized group higher than that of the nonrandomized group. The adjustment in Model D will fall short if there is an interaction between duration and study design, and Figure 1 suggests that this is the case. The problem here is that there is essentially no data to compare the coaching effects for randomized and nonrandomized studies once duration goes beyond 25 hours. To the extent that such a comparison can be controlled using just the main effect of duration, it is being made on the basis of extrapolating the regression line through imaginary data. These sorts of extrapolations become even more extreme as regression adjustments become multivariate (i.e., controlling for all the variables listed in Table 3). The specified model also assumes that the effect of coaching is linearly related to duration. However, one of the principal contributions of a review by Messick and Jungeblut (1981) was to demonstrate that the relationship is nonlinear, a finding consistent with Figure 1. All this helps explain why the predicted results of Model D are so at odds with what is actually observed in coaching studies.

The specifications of Becker's meta-analytic models raise issues germane to the use of meta-analysis in other contexts. I summarize three important ones below:

1. *Choice of variables included and excluded from the model*. Are the variables specified in meta-analytic regression models included for substantive or statistical concerns? In the example here, it seems clear that statistical, rather than substantive, concerns are the driving force behind model selection.
2. *Bias due to omitted variables and errors in variables*. It is highly likely that the estimated coefficients for meta-analytic regression models will be biased due to both omitted variables that confound the observed partial associations and variables that have been poorly measured. If the estimated coefficients are biased, interpretations based upon them become dubious.
3. *Interactions*. Meta-analytic regressions are typically specified without including interactions terms. For example, none of the models in Becker's meta-analysis take into account probable interactions between treatment and study characteristics. It is precisely these sorts of complicated interactions that make SAT coaching studies so difficult to summarize quantitatively.

Ultimately, the answers to these sorts of questions can only be settled by relying on the judgment of the researcher. It seems to me that we delude ourselves if we think that researcher subjectivity is any less of a threat to the findings from a meta-analysis than it is to the findings of the narrative literature review. In the next section, I provide examples of subjective decisions that influenced the findings from Becker's meta-analysis.

QUANTIFYING STUDIES IN META-ANALYSIS

Missing important details: The Federal Trade Commission (FTC) report. The findings from a meta-analysis depend upon how the meta-analyst decides to quantify studies under review. These decisions range from the way effect sizes are calculated, and sometimes imputed, to the way that treatment and design characteristics are coded. This process is subjective and mistakes are easily made. By way of illustration, I will discuss the FTC report. In 1976, the FTC initiated a study of commercial coaching companies. The FTC subpoenaed enrollment data for courses at three coaching companies in the New York area between 1975 and 1977. PSAT[9] and SAT scores for these students were subpoenaed from the College Board. FTC staff selected a control group of SAT test-takers at random from high schools in the same geographic areas. Demographic and academic background characteristics for coached and uncoached students were taken from the Student Descriptive Questionnaire, filled out by students before they take the SAT. Coaching at Company A consisted of a 10-week course with 4 hours of class per week split between preparation for the SAT-V and SAT-M. At Company B. coaching was shorter in duration, spanning 24 hours of classroom instruction on both sections of the test. Coaching at Company C was not analyzed by the FTC because of the small number of students involved.

The initial report on the FTC study was released as a staff memorandum by the FTC's Boston Regional Office (1978). Although this memo reported SAT-V and SAT-M coaching effects at Company A as large as 55 and 40 points, it was strongly criticized by the central administration of the FTC on the basis of flaws in the data analysis. The central administration suspected that the estimated coaching effects were biased. Coaching effects had been estimated by comparing coached and uncoached students that were, on average, not equivalent with respect to variables such as socioeconomic status and academic background.

The data were subsequently reanalyzed by the FTC's Bureau of Consumer Protection (1979), and this reanalysis was later published by Sesnowitz, Bernhardt, and Knain (1982). Coaching effects were reestimated for Companies A and B using linear regression models, where academic background, demographic characteristics, and test-taking experience were held constant. Coaching had statistically significant SAT-V and SAT-M effects of 30 points per section for Company A but small and statistically insignificant effects for Company B.

The Bureau of Consumer Protection's (BCP's) reanalysis of the FTC data was subject to further critique. ETS researchers joined the fray with two different statistical analyses (Rock 1980; Stroud 1980). Stroud (1980) used a larger set of covariates, made adjustments for missing data, and allowed for interaction effects between coaching and other variables. Nonetheless, Stroud's basic findings supported those of the BCP. Rock (1980) focused on the BCP finding that coaching by Company A was as effective for the SAT-V as it was for the SAT-M. Rock was skeptical because previous studies had suggested the SAT-M was more susceptible to coaching than the SAT-V (Dyer 1953; Dear 1958; Evans and Pike 1973). This seemed intuitively plausible because the SAT-M had closer ties to the typical high school curriculum. The BCP estimates had been made under the assumption that after controlling for covariates, the coaching effect could be calculated as the difference in rates of learning between coached and uncoached students. Rock posed a question. Suppose coached students were more motivated and academically able. Would they learn faster than uncoached students, even without coaching? If students who get coached are simply faster learners, the effect estimated in the FTC study would be too high. Rock's analysis considered students who took the PSAT once and the SAT twice but who received coaching after taking the SAT for the first time. After calculating PSAT-V to SAT-V gains, Rock found evidence that the SAT-V scores of the "to be coached" students were already growing at a significantly faster rate than those of control students. A similar pattern was not found for SAT-M scores. Taking these growth rates into account, Rock estimated Company A's SAT-V coaching effect as 17 points. No such adjustment was proposed for the SAT-M coaching effect.

The FTC investigation was a landmark event. It was by far the largest coaching study to be conducted, with total samples of 556 coached and 1,566 uncoached students. It was the first study to focus extensively on the effects of commercial coaching programs. It was the first case of multiple researchers employing a variety of statistical approaches with the hope of reducing bias in estimated effects. It was one of the first controlled studies to suggest sizeable effects from a coaching program (Company A).

These sorts of fascinating details are lost in Becker's meta-analysis. Moreover, some of the coding seems to be wrong. The FTC report is coded as unpublished, suggesting a report of lower quality that has "not undergone the review process typical of most academic journals" (Becker 1990, 397). But this report was subsequently published in an academic journal (Sesnowitz, Bernhardt, and Knain 1982). And it is the most peer-reviewed analysis in the history of SAT coaching reports. In addition, only data from Company A is included in Becker's meta-analysis.

Calculating effect sizes for meta-analytic regressions. In meta-analysis, the decision of how to compute comparable effect sizes across studies is often a subjective one. There are three key issues here. First, whether adjusted or unadjusted effect sizes should be computed. Second, how effect sizes should be computed when means and standard deviations are not reported in primary studies. Third, how effect sizes should be computed for studies with no control groups. Below I examine how these issues are dealt with in Becker's meta-analysis.

The effect sizes Becker calculates do not take into account any statistical adjustments. The effect is simply the mean difference in SAT score changes between coached and uncoached groups. But when comparisons between the two groups are confounded by omitted variables, researchers are likely to make statistical adjustments to their effect estimates in an attempt to take this source of bias into account.[10] In fact, three different SAT-V coaching effects were estimated for the FTC study of Company A. The FTC Boston Regional Office (1978) estimate was 55 points. The adjusted estimate from the FTC BCP (1979) and Stroud (1980) reanalysis was 30 points. The adjusted estimate in the reanalysis by Rock (1980) was 17 points. Which effect estimate should be used in the meta-analysis? Becker's analysis uses the unadjusted 55-point estimate, but a different meta-analyst might make an equally compelling case for using the 17-point estimate.

Becker's calculation of effect sizes requires that for each coaching study within a report, one has the mean and standard deviation of SAT scores for both coached and uncoached groups (recall Equations 1 and 2). A number of coaching studies do not report these sorts of descriptive statistics. In such cases, the descriptive statistics have been estimated by Becker. For example, in the coaching study by Kintisch (1979), standard deviations are not reported. They are estimated from the data as a function of the observed range of SAT scores and a test-retest correlation provided by ETS. A similar sort of strategy is adopted for other studies that fail to report the descriptive statistics necessary to calculate effect size measures. A different meta-analyst, of course, might employ a different strategy for estimating descriptive statistics.

Effect sizes must also be computed for studies that do not include control groups. In Becker's meta-analysis, there are five reports with no control group (Marron 1965; Pallone 1961; Coffman and Parry 1967; Coffin 1987; Johnson 1984). These reports provide 13 different measures of mean SAT score changes for coached students (g_{hi}^{C}), but no comparable measures for uncoached students (g_{hi}^{U}). It would seem that for 13 of the 70 studies under review, there is no available measure for coaching effect size. Instead, for these studies estimates of g_{hi}^{U} were imputed by Becker as a weighted average from all other uncontrolled coaching studies. It is worth noting that 11 of the 13 studies with no control groups derive from three reports conducted between 1960 and 1967 on special samples of students (Marron 1965; Pallone 1961; Coffman and Parry 1967):

- Pallone (1961) used a sample of boys from a private college preparatory high school in Washington, D.C., including a number of high school graduates completing a year of post-high-school study to prepare for entrance into U.S. government academies.
- Marron's (1965) sample consisted of male students at 10 well-known preparatory schools that specialized in preparing high school seniors and graduates for admission to service academies and selective colleges.
- Coffman and Parry (1967) used a small sample of students already enrolled in a public university.

Is the imputation of g_{hi}^{U} for these samples a plausible procedure? Consider two of the largest reports that serve as sources of data for the imputation. Alderman and Powers (1980) and FTC Boston Regional Office (1978) involved samples of both private and public, male and female high school students. So in effect, for the imputation of g_{hi}^{U} in the studies missing control groups, one is taking the score change in two SAT testings for Jennifer Smith, a 16-year-old student enrolled in a public New York high school in 1978, and substituting this for the unobserved score change of Alfred Buckley III, an 18-year-old student enrolled in a private Washington, D.C., prep school in 1959.

Coding predictor variables for meta-analytic regressions. The coding of variables that measure treatment characteristics and study quality is probably the most controversial stage of any meta-analysis. As Becker (1990) noted in describing the limitations of her meta-analysis,

> Many of the values of important predictors could not be determined from the reports of coaching studies. Duration was not explicitly reported in 9 of 23 reports. . . . 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. (P. 403)

This seems like a fair assessment. Coaching studies typically give sparse detail about the instructional nature of the treatment. But when the indicators are crude, the likelihood for coding errors is high. Is it sensible to use these sorts of indicators to build a statistical model?

Even aspects of study quality that appear straightforward to code may raise further questions about model specification. Coding a study with a dummy variable for use of randomization may not seem controversial. But many randomized studies lose subjects through attrition. Where is the dummy variable that makes this distinction? One can certainly distinguish between randomized, observational and uncontrolled studies, but should one ignore obvious and not so obvious distinctions within these groupings? It particular, the practice of grouping studies according to whether they are published is likely to be a weak indicator of quality. In the context of coaching studies, distinctions of quality on the basis of publication status are likely to be either meaningless, or, as in the FTC example, misleading. Most coaching studies are made public through peer-reviewed journals, institutional research reports, and doctoral dissertations. In each of these categories, it is relatively easy to find examples of both high- and low-quality study designs.

Consider two reports used in Becker's meta-analysis, one by Kintisch (1979) and one by Burke (1986). The former report was published in the *Journal of Reading*; the latter comes from a doctoral dissertation completed at Georgia State University. Both reports were similar in that they evaluated the effect on SAT-V scores of a school-based reading program offered over the course of a semester. The coding of design characteristics for the two reports by Becker is virtually identical. The only code that seemingly distinguishes the two study designs is that one was published and the other was not. Yet based on the respective study designs described by the authors, the study by Burke appears to be of much higher quality than the study by Kintisch. Burke analyzed separate samples of 11th- and 12th-grade students who had taken one semester of her reading program; Kintisch grouped together all 12th-grade students who had taken her elective course over a 3-year period. Burke took great pains in her study to demonstrate empirically that students in both coached and uncoached conditions were equally motivated to perform well on the SAT, making it less likely that the validity of her effect estimates is threatened by selection bias; Kintisch provided no such demonstration.

My point in this section is not so much to take Becker's meta-analysis to task for many subjective decisions with which I disagree but to note that these sorts of equivocal decisions come with the meta-analytic territory. They are to be expected when one seeks to quantify the nuances of studies into a workable set of continuous and categorical variables. One might expect to see similar questionable groupings and errors in study interpretation from a narrative review. But in the latter context, such decisions are likely to be discussed as part of the narrative, not buried within the machinery of the meta-analytic approach. It is certainly possible to salvage meta-analytic interpretations by better coding decisions and model specifications. Codings and model specifications are after all, adjustable. But first one must consider the statistical assumptions upon which the meta-analysis depends. These assumptions are not adjustable when they are violated.

ASSUMPTIONS OF THE META-ANALYTIC MODEL

The statistical foundation of meta-analytic models is seldom given sufficient attention.[11] This foundation requires one to accept a certain set of assumptions about the way the data underlying a meta-analysis have been generated. These assumptions are often at odds with what is known about the studies under review. It is helpful to put these assumptions into a grounded context, where one can contrast them with the data we actually observe. In the context of SAT coaching studies, the following assumptions are assumed to be true.

Within each coaching study (*i*),

I. SAT scores are independent within coached and uncoached student groups, and

II. SAT scores are independent across coached and uncoached student groups.

Across each report (*h*),

III. estimated coaching effects are independent.

Assumption I means that for any given study sample, how any one coached student performs on the SAT is unrelated to how any other coached student performs on the SAT. Likewise, the performance of any uncoached student is unrelated to the performance of any other uncoached student. Assumption II means that the performance of uncoached students on the SAT is unrelated to that of coached students. Assumption III maintains that the results from one report have no influence on the results from another report. One assumes further that within coached and uncoached groups, SAT scores have identical normal distributions with a common mean and variance. Recall that *X* and *Y* represent first and second testings on the SAT. For coached students,

$$X_{hij}^{C} \sim N\left(\mu_{hi}^{C}, \sigma_{hi}^{2}\right) \textit{ and } Y_{hij}^{C} \sim N\left(\nu_{hi}^{C}, \sigma_{hi}^{2}\right), \tag{4}$$

and for uncoached students,

$$X_{hij}^{U} \sim N\left(\mu_{hi}^{U}, \sigma_{hi}^{2}\right) \textit{ and } Y_{hij}^{U} \sim N\left(\nu_{hi}^{U}, \sigma_{hi}^{2}\right). \tag{5}$$

The variance term σ_{hi}^{2} is presumed to be constant across testings for both coached and uncoached students. The parameters $\mu_{hi}^{C}, \nu_{hi}^{C}, \mu_{hi}^{\mathrm{U}}, \nu_{hi}^{U}$, and σ_{hi}^{2} are all unobservable *population*-level quantities. But what, exactly, are the target populations? Within coaching studies, there are several possibilities, from the most general to the more specific:

1. Anyone in the United States who could take the SAT.
2. Anyone in the United States who did take the SAT.
3. High school students in the United States who plan to attend college and take the SAT.
4. High school seniors in private schools in the northeastern United States who planned to attend college and took the SAT twice between 1955 and 1974.

In Becker's meta-analysis, the target populations are not explicitly defined. None of the students in the coaching studies analyzed derive from probability samples drawn from any defined population.

Without random sampling, statistical inference leans heavily upon the assumptions of independence, represented above by I, II, and III.[12] How plausible are these assumptions in the coaching literature? Consider first Assumption I: SAT scores within treatment groups are independent. In many studies, coached students are exposed as a group to instruction from a single teacher (Dyer 1953; French 1955; Dear 1958; Kintisch 1979; Burke 1986; Zuman 1988). It follows that SAT scores are likely to be dependent across students. If students form study groups, that is another source of correlation. Now consider Assumption II: SAT scores across treatment groups are independent. The use of retired SAT exams and delayed treatment conditions in a number of randomized coaching studies make independence across coached and uncoached groups unlikely (Roberts and Openheim 1966; Alderman and Powers 1980; Zuman 1988). In these studies, some students were initially assigned to control groups that were tested twice with retired versions of the SAT. Surprisingly, mean SAT scores often *decreased* from one testing to the next for the control groups. Knowing that they were not taking an official SAT examination, it appears the control students may have been less motivated to do their best on the test relative to coached students taking two official administration of the SAT. In this scenario, knowledge by the control group that the treatment group was both getting coached and taking an official administration of the SAT influenced their test performance.

In assessing the plausibility of Assumption III (independence across reports),

> Investigators are trained in similar ways, read the same papers, talk to one another, write proposals for funding to the same agencies, and publish the findings after peer review. Earlier studies beget later studies, just as each generation of Ph.D. students train the next. After the first few million dollars are committed, granting agencies develop agendas of their own, which investigators learn to accommodate. Meta-analytic summaries of past work further channel this effort. There is, in short, a web of social dependence inherent in all scientific research. Does social dependence compromise statistical independence? Only if you think that investigators' expectations, attitudes, preferences and motivations affect the written word—and never forget those peer reviewers! (Berk and Freedman 2003, 12)

The "web of social dependence" Berk and Freedman (2003) described is especially relevant in the context of SAT coaching reports. Many of these reports built explicitly upon the designs of previous reports (Dyer 1953; Dear 1958; French 1955). The care given to the definition of the coaching treatment (Alderman and Powers 1980) also changed after a spate of coaching reviews were published during the 1980s, reviews that were themselves prompted by key reports and reviews (Pallone 1961; Marron 1965; Evans and Pike 1973; FTC Boston Regional Office 1978; Slack and Porter 1980; Jackson 1980).

Certain adjustments to meta-analytic assumptions are sometimes made. Becker's meta-analytic regressions derive from a fixed effects model. One adjustment to the model would be to assume that effect size estimates derive from some population of random or mixed effects. In a random-effects model, each coaching effect from a study is assumed to be drawn as a random sample from some population of coaching effects. Now the task is to estimate the mean of these random effects. In the mixed-effects formulation, each coaching effect has both a fixed component and an intrinsic random error component. Had a random- or mixed-effects model been employed for the SAT coaching meta-analysis, one would need to consider the population of reports to which meta-analytic inferences are to be drawn. An immediate problem emerges, because with the exception of a few reports that Becker was unable to obtain, the reports under review *are* the full population of SAT coaching studies. One may invent a superpopulation of hypothetical studies that *could* have been conducted, but the assumption that the reports under review are a random sample from this superpopulation is a tough pill to swallow.

Many statements in meta-analysis are couched in the inferential language of significance testing. One must be clear about the interpretation of p values associated with regression coefficients in a meta-analysis. Say one runs a meta-analytic regression and estimate the parameter $\hat{b}$ with an associated p value of .05. Assuming the population parameter of interest, b, is really zero, if a limitless number of independent random samples were drawn from the same population, one would expect to estimate a parameter as large as $\hat{b}$ 5% of the time. This sort of thought experiment does not correspond to the way data are actually being generated both within and across studies under review by the meta-analyst. Samples of studies are usually best characterized as the full population; samples within studies are typically selected as a function of convenience. As Berk and Freedman (2003) wrote,

> The moment that conventional statistical inferences are made from convenience samples, substantive assumptions are made about how the social world operates. Conventional statistical inferences (e.g. formulas for the standard error of the mean, *t*-tests, etc.) depend on the assumption of random sampling. This is not a matter of debate or opinion; it is a matter of mathematical necessity. When applied to convenience samples, the random sampling assumption is not a mere technicality or a minor revision on the periphery; the assumption becomes an integral part of the theory. (P. 2)

What are the consequences when independence assumptions are violated? Using simulated data, Berk and Freedman (2003) have shown that even very simple forms of positive dependence can lead to standard errors that are 50% larger than those that would calculated using conventional formulas. In real data sets, more complex forms of dependence are just as likely, with potential consequences for standard errors that may be considerably worse. If estimated standard errors are wrong, then inferences based upon *p* values and confidence intervals will be misleading. The reporting of confidence intervals and *p* values for synthesized effects (common practice in almost all meta-analyses) provide an illusory sense of statistical authority.

If one takes these sorts of criticisms seriously, the inferential use of meta-analysis may be doomed from the start, regardless of how well one quantifies studies and specifies one's model. Indeed, Gene Glass (2000), who first coined the term meta-analysis, seemed to come to the same conclusion when he wrote that "inferential statistics has little role to play in meta-analysis" (p. 14).

COMPARING BECKER'S CONCLUSIONS TO THOSE REACHED BY A NARRATIVE REVIEW

The principal conclusions reached by Becker in her SAT coaching review can, I believe, be fairly represented as follows:

1. Summaries of coaching effects should be conditional on whether
a. the study was published and
b. involved a control group.
2. The association between coaching duration and coaching effect is confounded by study design.
3. Regardless of study design, coaching effects for the SAT-M are consistently larger than those for the SAT-V.
4. A number of unusually high coaching effect estimates can be found among studies with no control groups, and among studies that are unpublished, even after holding constant other design characteristics.
5. The best summary of the overall coaching effect comes from published studies with control groups: 19 points on the SAT-M, 8 points on the SAT-V.

A reasonable question to ask is whether a narrative review (cf. Briggs 2002b) of this same set of meta-analyzed studies would come to different conclusions about the effectiveness of SAT coaching. In addition, one might ask whether these conclusions seem to be supported after reviewing the new reports that have appeared since Becker's meta-analysis.

Conclusion 1a, that coaching effect summaries should be conditional on whether a study involves a control group, was well established in prior reviews by Pike (1978), Messick (1980), Messick and Jungeblut (1981), Cole (1982), and Bond (1989). In fact, these reviews typically grouped coaching effect summaries by those studies with no control group, those with a control group assigned randomly, those with a matched control group, and those with nonequivalent control groups. With respect to Conclusion 1b, I have already argued that the categorization of studies by whether they have been published should not constitute a primary grouping for coaching effect summaries. The aim of the reviewer should be to evaluate the quality of all available studies, published or not. Depending on the criteria chosen for evaluation, some unpublished studies may be considered high quality, just as some published studies may be considered low quality. I would suggest that the best grouping of coaching effects is by the primary categories of study design (no control, observational control, randomized control) and mode of coaching delivery (school-based, commercial-based, computer-based). The full collection of coaching reports categorized by these criteria are summarized in Table 6. What becomes clear is that although the total number of coaching reports, 36, is sizeable, once reports are grouped by the primary categories of coaching mode and study design, the average number of reports per cell is rather small. The typical coaching study evaluates commercial or school-based coaching with an observational design.

Conclusion 2, that the association between coaching duration and effect is confounded by study design, was an important contribution of Becker's review. However, this conclusion could just as easily be reached without conducting a meta-analysis. Messick and Jungeblut (1981) analyzed this relationship by calculating the rank correlation between program hours and coaching effect by test section. In estimating Spearman rank correlations rather than Pearson product moment correlations, less weight is placed upon the specific point estimates of coaching effects. This mitigates the fact that the point estimates may well be biased to some extent. Under the rank correlation, studies with large and small effects have less influence, as the set of study effects are compared only in an ordinal sense. Messick and Jungeblut found strong correlations of .77 and .71 between program duration and effect for 19 and 14 SAT-V and SAT-M coaching studies, respectively.

In Table 7 I have replicated the Messick and Jungeblut (1981) analysis with three different collections of coaching studies. The first collection of studies is identical to those used by Messick and Jungeblut in their review. My results using these studies should be identical to those found by Messick

TABLE 6: Studies by Coaching Mode and Design

	Methodological Design		
Coaching Type	*Randomized Control*	*Observational Control*	*No Control*
School-based	Roberts and Openheim (1966) Evans and Pike (1973) Alderman and Powers (1980) Shaw (1992)	Dyer (1953) French (1955) Dear (1958) Keefauver (1976) Kintisch (1979) Johnson (San Francisco site) (1984)[a] Burke (1986) Reynolds and Oberman (1987) Harvey (1988) Wing, Childs, and Maxwell (1989) Schroeder (1992) Wrinkle (1996)	Pallone (1961) Marron (1965) Johnson (Atlanta, New York sites) (1984)[a]

Commercial-based		Frankel (1960)	Kaplan (2002)
		Whitla (1962)	
		Federal Trade Commission study and reanalyses	
		Boston Regional Office (1978)	
		Bureau of Consumer Protection (1979)	
		Rock (1980)	
		Stroud (1980)	
		Sesnowitz, Bernhardt, and Knain (1982)	
		Fraker (1987)	
		Whitla (1988)	
		Zuman (1988)[a]	
		Snedecor (1989)	
		Smyth (1989)	
		Smyth (1990)	
		Powers and Rock (1999)	
		Briggs (2001)	
Computer-based	Hopmeier (1984)		Coffin (1987)[a]
	Laschewer (1985)		
	Curran (1988)		
	Holmes and Keffer (1995)		
	McClain (1999)		

a. Design intent of these studies (randomized experimental) compromised by substantial sample attrition.

TABLE 7: Coaching Duration by SAT Coaching Effect Estimate

	Verbal SAT		
	Messick and Jungeblut (1981)[a]	*Messick and Jungeblut (1981) Updated*[b]	*Messick and Jungeblut (1981) Updated Subset*[c]
Number of estimates	19	30	24
Rank correlation	.712	.459	.312
	Math SAT		
	Messick and Jungeblut (1981)[d]	*Messick and Jungeblut (1981) Updated*[e]	*Messick and Jungeblut (1981) Updated Subset*[f]
Number of estimates	14	25	17
Rank correlation	.711	.481	.408

a. Basis for correlations: Dyer (1953), French (1955) (two program estimates), Dear (1958), Frankel (1960), Whitla (1962), Alderman and Powers (1980) (five program estimates), Pallone (1961) (two program estimates), Marron (1963) (four program estimates), FTC (1979) (two program estimates).
b. Basis for correlations: All studies and program estimates in Messick and Jungeblut plus Kintisch (1979), Hopmeier (1984), Johnson (1984), Laschewer (1985), Burke (1986), Zuman (1988) (two program estimates), Shaw (1992), Holmes and Keffer (1995), Wrinkle (1996), Powers and Rock (1999).
c. Basis for correlations: Excludes all program estimates from uncontrolled studies: Pallone (1961), Marron (1963).
d. Basis for correlations: Dyer (1953), French (1955), Dear (1958) (two program estimates), Frankel (1960), Whitla (1962), Evans and Pike (1973) (three program estimates), Marron (1963) (three program estimates), FTC (1979) (two program estimates).
e. Basis for correlations: All studies and program estimates in Note d. plus Hopmeier (1984), Johnson (1984), Laschewer (1985), Schroeder (1988), Schroeder (1992), Zuman (1988) (two program estimates), Shaw (1992), Powers and Rock (1999), Kaplan (2002) (two program estimates).
f. Basis for correlations: Excludes all program estimates from uncontrolled studies: Marron (1963), Kaplan (2002).

and Jungeblut. The second collection of studies adds to the first collection all new studies with relevant data conducted since the Messick and Jungeblut review. The third collection of studies considers the same set as the second collection but excludes those studies that lacked a control group. These results suggest two conclusions. First, it does not appear that coaching duration and effect have a strong linear association. For both sections of the test, the rank correlation drops by about 30% to 35% when new studies not reviewed by Messick and Jungeblut are included in the calculation. Second, it does appear that study design confounds the association between duration and effect. When new studies are included in the analysis, but uncontrolled studies are excluded from the calculation, the rank correlations between duration and effect decrease.

Conclusion 3, that coaching effects for the SAT-M are consistently larger than those for the SAT-V, regardless of study design, would almost certainly not have been reached through a narrative review. There are numerous examples of coaching interventions that produced larger effects on the SAT-V than on the SAT-M in reports released prior to Becker's meta-analysis. Hence consistency, irrespective of study design, would be tough to establish. The 14 new coaching reports not included in Becker's meta-analysis do lend support to the notion that on average, the SAT-M is more coachable than the SAT-V. This conclusion is best supported with respect to the estimated effects for commercial coaching programs with observational designs (cf. Fraker 1987; Whitla 1988; Snedecor 1989; Smyth 1990; Powers and Rock 1999; Briggs 2001). The evidence is less conclusive for studies with school-based coaching programs and studies with no control groups.

The first part of Conclusion 4, that a number of unusually high coaching effect estimates can be found among studies with no control groups, and among studies that are unpublished, is essentially a restatement of the findings first emphasized by Slack and Porter (1980) and then later discussed extensively in all subsequent reviews. Indeed, this seems to be the primary basis for the debate over the effectiveness of SAT coaching. For a recent example of this debate renewed, see Kaplan (2002) and Briggs (2002a). The second part of Conclusion 4, that unusually high effect estimates are found even after holding constant other design characteristics, depends on the assumption that it is sensible to "hold constant" with a meta-analytic regression. I have suggested that there is good reason to be skeptical about this.

Conclusion 5 is the most widely cited finding from Becker's SAT coaching meta-analysis: The best summary of the overall coaching effect is 19 points on the SAT-M, 8 points on the SAT-V. To arrive at this conclusion, Becker first separated coaching reports into those that were published and unpublished. After removing studies without control groups from the sample of published studies, Becker calculates a homogeneous estimate for the effect of coaching, without the need to control for either coaching program characteristics, or design characteristics. For such studies, the fixed effect meta-analytic model is

$$\hat{\Delta} = 8.8 + 6.9(\text{SAT-M}).$$

In words, the best meta-analytic model for synthesizing the effects of coaching should simply take the weighted average of SAT-V and SAT-M coaching study effect estimates, but only for those studies that have been published and involve control groups.

Are these results enlightening? For anyone following the debate over coaching effectiveness since the early 1980s, the answer should be no. The results of unpublished studies, and studies with no control groups, typically suggest that coaching might produce substantial effects on SAT performance. It is precisely the ambiguous interpretation of these studies that has fueled much of the controversy over coaching effectiveness. So it should come as little surprise that once one removes such studies from the meta-analysis, the model returns homogeneous and small estimates of coaching effects.

In addition, because the largest studies from the largest report in the SAT coaching literature were miscoded in the meta-analysis, the conclusion about the overall size of coaching effects was almost certainly wrong. Given the fairly large effects found in the FTC report, it is likely that had the results for Company A been correctly included in the meta-analytic evidence base as deriving from a published source, the weighted averages for the overall coaching effects would be higher.

When new literature on SAT coaching effects is added to the mix, the best guess for the overall effect of commercial coaching programs may be very close to the total effect of about 30 points suggested in the meta-analysis. I base this conclusion primarily on the two largest studies of commercial coaching using nationally representative samples (Powers and Rock 1999; Briggs 2001). However, it seems worth noting that there are also a number of studies that suggest coaching may be more effective for certain types of students (cf. Briggs 2004b) and certain types of programs (cf. Burke 1986; Schroeder 1992; Kaplan 2002). Such findings may indicate a direction for further research.

In summary, it seems to me that everything accomplished by Becker's meta-analysis could have been accomplished as well or better with a narrative review. The soundness of the conclusions Becker reaches in her meta-analysis rely upon her judgments as a researcher, not upon the rigor or objectivity of the meta-analytic approach. Like a narrative review, these judgments are often debatable. Unlike a narrative review, these judgments are more easily disguised by equations and parameter estimates.

CONCLUSION

The traditional alternative to meta-analysis is the oft-maligned narrative literature review. A collection of studies is gathered, read, and evaluated. The reviewer starts with a set of questions and searches for trends and patterns that help answer these questions. Finally, the reviewer reports on the trends and patterns that, to his or her eye, emerge. This is unquestionably very hard work, and it requires both science and art. For two examples of what I consider exemplary narrative reviews (also within the domain of SAT coaching studies), I would encourage readers to consult Messick (1980) or Messick and Jungeblut (1981). Nuances are explored and insights are provided without the need for a single meta-analytic regression.

The process of conducting a narrative review might be criticized as overly subjective. One of my aims here has been to show that the meta-analysis shares the same inherent subjectivity. The problem with narrative reviews of poor quality is typically not that they are biased but that they fail to be systematic. It seems clear that an unsystematic review will be of poor quality regardless of the specific methodological approach. For a review to be compelling, the criteria for including and evaluating research reports should be made explicit. To the extent that this is a central feature of the meta-analysis, it is a feature that should be duplicated in a narrative review.[13]

Meta-analyses can be used for both descriptive and inferential purposes. The case study presented here suggests that there are good reasons to be cautious even when a meta-analysis is used solely for descriptive purposes. Meta-analytic regressions may imply interpretations that are not warranted when one compares them to the observed findings of the underlying studies. One reason for this is that there is no grounded theory to guide the specification of meta-analytic regression models. Another reason is that the process of quantifying studies is problematic. Important details may get lost, just as subjective decisions will almost invariably lead to coding errors. The use of meta-analysis for inferential purposes rests upon a very uncertain foundation in social science applications. When certain assumptions do not hold, the reporting of confidence intervals and *p* values for synthesized effects will have dubious interpretations. In the case study considered here, Becker's meta-analytic regressions fare poorly when used to predict the outcomes of new coaching studies. It would be interesting and informative (though outside the scope of this study) to apply the same test to other well-established meta-analyses.

There may be very specific research contexts where the use of meta-analysis is reasonable. Such contexts would seemingly require a set of studies with

- randomized experimental designs,
- carefully defined treatments, and
- homogeneous samples from a large but well-defined population of interest.

Petitti (2000) suggested that this sort of context sometimes exists in medical research. In educational research, the sort of context described above is exceedingly rare. Outside of this context, the potential for meta-analysis to obfuscate as much or even more than it enlightens is high.

No single methodological framework can ensure the validity of conclusions drawn from a quantitative literature review. As Cronbach (1982, 108) noted some time ago, commenting on a related context, “Validity depends not only on the data collection and analysis but also on the way a conclusion is stated and communicated. Validity is subjective rather than objective: the plausibility of the conclusion is what counts. And plausibility, to twist a cliché, lies in the ear of the beholder.”

APPENDIX A
SAT Coaching Reports, 1953-2001

	Sample Size[a] (Coached/Total)						
Study	*SAT-V*	*SAT-M*	*Grade Level*	*School Type*	*Location*	*Year(s) Tested*	*SES of Sample[b]*
Uncontrolled studies							
School-based coaching							
Pallone (1960)	100	NA	Precollege	1 private (all male)	D.C.	1959	High
Marron (1965)	714	715	11th, 12th	10 private (all male)	D.C.	1962	High
Johnson (Atlanta and NYC sites) (1984)	117	116	11th	Multiple public (all Black, urban)	NY, GA	1983-1994	Low
Commercial coaching							
Kaplan (2002)	NA	18	12th	Multiple public and private	CT	1999-2000	High
Computer-based coaching							
Coffin (1987)	18	18	11th, 12th	1 public (urban)	MA	1986-1987	Low
Observational studies							
School-based coaching							
Dyer (1953)	225/418	225/418	12th	2 private (all male)	NR	1951-1952	High
French (1955)	161/319	161/319	12th	3 public	MI, MA	1954	High
Dear (1958)	60/586	60/586	12th	Multiple public and private	NJ, NY, PA	1956-1957	High
Lass (1961, cited in Pike 1978)	38/120	38/120					Mixed
Keefauver (1976)	16/41	16/41		M	M	M	High

(continued)

APPENDIX A (continued)

	Sample Size[a] (Coached/Total)						
Study	*SAT-V*	*SAT-M*	*Grade Level*	*School Type*	*Location*	*Year(s) Tested*	*SES of Sample*[b]
Kintisch (1979)	38/76	NA	12th	1 public (suburban)	PA	1976-1978	NR
Burke (1986)	50/100	50/100	11th, 12th	1 public (suburban)	GA	1984-1985	Mixed
Reynolds and Oberman (1987)	93/140	93/140		M	M	M	High
Harvey (1988)	NA	21/54	11th	2 public (urban)	GA	1987	Mixed
Wing, Childs, and Maxwell (1989)	173/253	173/253	11th	Statewide	NC	1986, 1987	High
Schroeder (1992)	NA	59/95	NR	1 public (urban)	NY	1991-1992	High
Wrinkle (1996)	18/36	NA	9th, 10th, 11th	1 public (suburban)	TX	NR	High
Commercial coaching							
Frankel (1960)	45/90	45/90	12th	1 public (urban)	NY	1958	High
Whitla (1962)	52/104	50/100	11th	Multiple public and private	MA	1959	High
Federal Trade Commission Boston Regional Office (1978)/ Bureau of Consumer Protection (1979)	556/2,122	556/2,122	11th, 12th	Multiple public and private (urban)	NY	1974-1977	Mixed
Fraker (1987)	19/138	19/138	12th	1 private	MA	1986	High
Whitla (1988)	341/1,558	341/1,558	12th	Multiple public and private	USA	1986-1987	High

Zuman (high-SES sample) (1988)	21/55	21/55	11th	Multiple public (urban)	NY	1985-1986	High
Smyth (1989)	200/438	200/438	12th	8 private (suburban)	MD, D.C.	1987-1988	High
Snedecor (1989)	264/535	264/535	12th	10 public and private	PA	1988-1989	High
Smyth (1990)	631/1,132	631/1,132	12th	14 private (suburban)	MD, NJ	1989	High
Powers and Rock (1999)	427/2,086	427/2,086	11th, 12th	Multiple public and private	USA	1995-1996	Mixed
Briggs (2001)	503/3,144	503/3,144	11th, 12th	Multiple public and private	USA	1991-1992	Mixed
Randomized studies							
School-based coaching							
Roberts and Openheim (1966)	154/265	188/310	12th	18 public (all Black, urban, and rural)	TN	1965	Low
Evans and Pike (1973)	NA	288/417	11th	12 public (urban and suburban)	NJ, OH, PA	1970-1971	Mixed
Alderman and Powers (1980)	239/559	NA	11th	8 public and private	7 New England states	1977-1978	Mixed
Johnson (San Francisco site) (1984)	23/35	23/35	11th	Multiple public (all Black, urban)	CA	1983-1994	Low
Shaw (1992)	61/122	61/122	12th	3 public (suburban)	CA	1988	Mixed
Commercial coaching							
Zuman (low-SES sample) (1988)	16/33	16/33	11th	Multiple public (urban)	NY	1985-1986	Low

(continued)

APPENDIX A (continued)

	Sample Size[a] *(Coached/Total)*						
Study	*SAT-V*	*SAT-M*	*Grade Level*	*School Type*	*Location*	*Year(s) Tested*	*SES of Sample*[b]
Computer-based coaching							
Hopmeier (1984)	42/71	61/93	9th, 10th, 11th	1 public (suburban)	FL	NR	Mixed
Laschewer (1985)	13/27	13/27	11th	1 private (suburban Catholic)	NY	NR	Mixed
Curran (1988)	204/408	204/408	11th	4 private (Catholic)	MA	1986-1987	Mixed
Holmes and Keffer (1995)	28/58	NA	12th	1 public (rural)	GA	1990	Mixed
McClain (1999)	40/60	40/60	12th	Public (suburban)	MD	1998	Low

NOTE: NA = not applicable; NR = not reported; M = missing data. Reports in italics are those reviewed in Becker's (1990) meta-analysis.

a. Samples presented here are summed across all coached and uncoached subsamples considered in given study unless otherwise noted.

b. Approximate socioeconomic status (parental income, education, occupation) of sample on average according to author.

APPENDIX B
Coding Treatment Characteristics of New Studies

	Grand		*Control*												
Study	*Mean*	*SAT-M*	*Group*	*D*	*VI*	*MI*	*AI*	*IP*	*TP*	*TS*	*OA*	*HW*	*CI*	*WC*	*AC*
Hopmeier	1	1	1	3.5	1	1	1	1	0	1	0	0	1	0	0
Fraker	1	1	1	**15**	1	1	0	1	1	1	0	0	0	0	0
Harvey	1	1	1	4	0	1	1	1	0	1	0	0	0	0	1
Whitla	1	1	1	**15**	1	1	1	1	1	1	0	0	0	0	0
Snedecor	1	1	1	**15**	1	1	1	1	1	1	0	0	0	0	0
Wing, Childs, and Maxwell	1	1	1	**15**	1	1	1	1	1	1	1	0	0	0	0
Smyth	1	1	1	**15**	1	1	1	1	1	1	0	0	0	0	0
Shaw	1	1	1	4	1	1	0	0	0	1	0	0	0	0	0
Schroeder	1	1	1	16	0	1	2	1	1	1	1	0	0	0	0
Holmes and Keffer	1	0	1	8	1	0	1	0	0	0	0	0	1	0	0
Wrinkle	1	0	1	68	1	0	2	1	1	1	0	0	0	0	0
Powers and Rock	1	1	1	**15**	1	1	1	1	1	1	0	0	0	0	0
Briggs	1	1	1	**15**	1	1	1	1	1	1	0	0	0	0	0
Kaplan Year 1	1	1	0	30	0	1	2	1	1	1	1	1	0	0	0
Kaplan Year 2	1	1	0	30	0	1	2	1	1	1	1	1	0	0	0

NOTE: D = duration of coaching in hours (bold values have been imputed as in Becker's [1990] review), VI = verbal instruction, MI = math instruction, AI = alpha instruction, IP = item practice, TP = test practice, TS = test-taking skills, OA = other activities, WC = wait-list control, AC = alternative control.

APPENDIX C
Coding Design Characteristics of New Studies

Study	*Year*	*Pub*	*Match*	*Rand*	*ETS*	*Sel*	*Vol*
Hopmeier	82	0	0	1	0	1	2
Fraker	87	0	0	0	0	2	2
Harvey	88	0	0	0	0	1	2
Whitla	88	1	0	0	0	2	2
Snedecor	89	1	0	0	0	2	2
Wing, Childs, and Maxwell	89	0	0	0	0	2	2
Smyth	90	1	0	0	0	2	2
Shaw	92	0	0	1	0	1	2
Schroeder	92	0	0	0	0	2	2
Holmes and Keffer	95	1	0	1	0	2	2
Wrinkle	96	0	1	0	0	2	2
Powers and Rock	99	1	0	0	1	1	2
Briggs	101	1	0	0	0	1	2
Kaplan Year 1	101	1	0	0	0	2	2
Kaplan Year 2	101	1	0	0	0	2	2

NOTE: D = duration of coaching in hours (bold values have been imputed as in Becker's [1990] review); VI = presence of verbal instruction; Year = publication year; Pub = published in peer-reviewed journal; Match = use of matching; Rand = randomized design; ETS = sponsored by Educational Testing Service; Sel = selectivity of sample, Vol = voluntariness of sample.

NOTES

1. There are two exceptions. I was unable to track down two SAT coaching reports: Keefauver (1976) and Reynolds and Oberman (1987). The former is a doctoral dissertation and the latter is a conference paper.

2. As of 1994, the SAT became the SAT I: Reasoning Test. For simplicity, I use the term SAT throughout.

3. I am skipping one step. Becker makes an adjustment for bias in g_{hi}^{C} and g_{hi}^{U} (for details, see Becker 1988). The adjustment has no bearing on the issues being raised here.

4. I will not use the report by McClain (1999) in the subsequent analysis because the paper does not provide estimates for each section of the SAT. I also exclude the report by Smyth (1989) because the sample from that report appears to overlap with that used in his more detailed 1990 report.

5. Becker also specifies meta-analytic regressions using subsamples of SAT coaching reports, for example, modeling the estimated effects only for published reports, modeling the estimated effects only for unpublished reports. The root mean square error (RMSE) of these models tells a similar story to the one summarized in Table 2.

6. Defined by Bond (1989) as instruction geared toward the latent domain represented by the test score, that is, the composite of underlying knowledge and reasoning ability developed over a long period of time. Bond contrasted alpha instruction with beta instruction, which he defined as instruction intended to improve general and specific test wiseness.

7. These dummy variables have obvious overlap with variables intended to control for design characteristics. It is unclear why they have been included as part of Model C rather than Model D.

8. The latter finding is influenced by the seven observations drawn from the same coaching report by Marron (1965).

9. The PSAT is essentially a pretest of the SAT, administered to most high school students in the 10th grade. The PSAT has the same format as the SAT, and performance on the PSAT is strongly correlated with performance on the SAT.

10. In addition to the Federal Trade Commission report, these sorts of statistical adjustments were made in the reports by Dyer (1953), French (1955), Dear (1958), Zuman (1988), and Alderman and Powers (1980).

11. For a more general and detailed presentation of this foundation, see Hedges and Olkin (1985).

12. For a more general presentation of this issue, see Berk (2004, chap. 4).

13. This is hardly a new idea. See, for example, Bonde and Magistrate (1987). The importance of conducting literature reviews that are systematic receives strong emphasis.

REFERENCES

Alderman, D. L., and D. E. Powers. 1980. The effects of special preparation on SAT-verbal scores. *American Educational Research Journal* 17:239-53.

Anastasi, A. 1981. Coaching, test sophistication, and developed abilities. *American Psychologist* 36 (10): 1086-93.

Becker, B. J. 1988. Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology* 41:257-78.

———. 1990. Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research* 60 (3): 373-417.

Berk, R. A. 2004. *Regression analysis: A constructive critique.* Thousand Oaks, CA: Sage.

Berk, R. A., and D. A. Freedman. 2003. Statistical assumptions as empirical commitments. In *Law, punishment, and social control: Essays in honor of Sheldon Messinger*, 2nd ed., ed. T. G. Blomberg and S. Cohen, 235-54. http://www.stat.berkeley.edu/census/berk2.pdf.

Bond, L. 1989. The effects of special preparation on measures of scholastic ability. In *Educational measurement*, ed. R. L. Linn, 429-44. New York: American Council on Education/ Macmillan.

Bonde, L. A., and A. S. Magistrate. 1987. *Writer's guide: Psychology.* Lexington, MA: D. C. Heath.

Briggs, D. C. 2001. The effect of admissions test preparation: Evidence from NELS:88. *Chance* 14 (1): 10-18.

———. 2002a. Comment: Jack Kaplan's A new study of SAT coaching. *Chance* 15 (1): 7-8.

———. 2002b. SAT coaching, bias and causal inference. Ph.D. diss., University of California, Berkeley.

———. 2004a. Causal inference and the Heckman model. *Journal of Educational and Behavioral Statistics* 29 (4): 397-420.

———. 2004b. Evaluating SAT coaching: Gains, effects and self-selection. In *Rethinking the SAT: Perspectives based on the November 2001 conference at the University of California, Santa Barbara*, ed. by R. Zwick, 217-34. New York: RoutledgeFalmer.

Burke, K. B. 1986. A model reading course and its effect 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. Ph.D. diss., Georgia State University, Atlanta.

Coffin, G. C. 1987. Computer as a tool in SAT preparation. Paper presented at the Florida Instructional Computing Conference, Orlando, FL.

Coffman, W. E., and M. E. Parry. 1967. Effects of an accelerated reading course on SAT-V scores. *Personnel and Guidance Journal* 46:292-96.

Cole, N. 1982. The implications of coaching for ability testing. In *Ability testing: Uses, consequences, and controversies. Part II: documentation section*, ed. A. K. Wigdor and W. R. Gardner. Washington, DC: National Academy Press.

Curran, R. G. 1988. The effectiveness of computerized coaching for the Preliminary Scholastic Aptitude Test (PSAT/NMSQT) and the Scholastic Aptitude Test (SAT). Ph.D. diss., Boston University.

Cronbach, L. J. 1982. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Dear, R. E. 1958. *The effect of intensive coaching on SAT scores*. Princeton, NJ: Educational Testing Service.

DerSimonian, R., and N. M. Laird. 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* 19:331-35.

Evans, F., and L. Pike. 1973. The effects of instruction for three mathematics item formats. *Journal of Educational Measurement* 10 (4): 257-72.

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*. Boston: Federal Trade Commission, Boston Regional Office.

Federal Trade Commission, Bureau of Consumer Protection. 1979. *Effects of coaching standardized admission examinations: Revised statistical analyses of data gathered by the Boston Regional Office of the Federal Trade Commission*. Washington, DC: Federal Trade Commission, Bureau of Consumer Protection.

Fraker, G. A. 1987. *The Princeton Review* reviewed. In *The Newsletter*. Deerfield, MA: Deerfield Academy.

Frankel, E. 1960. Effects of growth, practice, and coaching on Scholastic Aptitude Test scores. *Personnel and Guidance Journal* 38:713-19.

French, J. W. 1955. *The coach ability of the SAT in public schools*. Princeton, NJ: Educational Testing Service.

Glass, G. V. 2000. Meta-analysis at 25. http://glass.ed.asu.edu/gene/papers/meta25.html.

Glass, G. V, B. McGaw, and M. L. Smith. 1981. *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Harvey, K. S. 1988. Videotaped versus live instruction as a coaching method for the mathematics portion of the scholastic aptitude test. Ph.D. diss., University of Georgia, Athens.

Hedges, L. V. 1990. Directions for future methodology. In *The future of meta-analysis*, ed. K. W. Wachter and M. L. Straf, 11-26. New York: Russell Sage Foundation.

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

Holmes, C. T., and R. Keffer. 1995. A computerized method to teach Latin and Greek root words: Effect on verbal SAT scores. *Journal of Educational Research* 89 (1): 47-50.

Hopmeier, G. H. 1984. The effectiveness of computerized coaching for Scholastic Aptitude Test in individual and group modes. Ph.D. diss., Florida State University, Tallahassee.

Jackson, R. 1980. The Scholastic Aptitude Test: A response to Slack and Porter's critical appraisal. *Harvard Educational Review* 50 (3): 382-91.

Johnson, S. T. 1984. *Preparing Black students for the SAT—Does it make a difference?* An evaluation report of the NAACP Test Preparation Project. New York: National Association for the Advancement for Colored People.

Kaplan, J. 2002. A new study of SAT coaching. *Chance* 14 (4): 1-6.

Keefauver, L. W. 1976. The effects of a program of coaching on Scholastic Aptitude Test scores of high school seniors tested as juniors. Ph.D. diss., University of Tennessee at Knoxville.

Kintisch, L. S. 1979. Classroom techniques for improving Scholastic Aptitude Test scores. *Journal of Reading* 22:416-19.

Kulik, C.-L., and J. A. Kulik. 1988. Meta-analysis: Historical origins and contemporary practice. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kulik, J. A., R. L. Bangert-Drowns, and C.-L. Kulik. 1984. Effectiveness of coaching for aptitude tests. *Psychological Bulletin* 95:179-88.

Laschewer, A. 1985. The effect of computer assisted instruction as a coaching technique for the scholastic aptitude test preparation of high school juniors. Ph.D. diss., Hofstra University, Hempstead, NY.

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, NY: Research Division, Office of the Director of Admissions and Registrar, United States Military Academy.

McClain, B. 1999. The impact of computer-assisted coaching on the elevation of twelfth-grade students' SAT scores. Ph.D. diss., Morgan State University, Baltimore.

Messick, S. 1980. *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC*. Princeton, NJ: Educational Testing Service.

———. 1982. Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *The Educational Psychologist* 17 (2): 67-91.

Messick, S., and A. Jungeblut. 1981. Time and method in coaching for the SAT. *Psychological Bulletin* 89:191-216.

Oakes, M. 1986. *Statistical inference: A commentary for the social and behavioral sciences*. New York: John Wiley.

Pallone, N. J. 1961. Effects of short- and long-term developmental reading courses upon the S.A.T. verbal scores. *Personnel and Guidance Journal* 39:654-57.

Petitti, D. B. 2000. *Meta-analysis, decision analysis and cost-effectiveness analysis*. 2nd ed. New York: Oxford University Press.

Pike, L. W. 1978. *Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with research recommendations*. Princeton, NJ: Educational Testing Service.

Powers, D. E. 1993. Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice* 12 (2): 24-39.

Powers, D. E., and D. A. Rock. 1999. Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement* 36 (2): 93-118.

Reynolds, A., and G. O. Oberman. 1987. 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, April.

Roberts, S. O., and D. B. Openheim. 1966. *The effect of special instruction upon test performance of high school students in Tennessee*. Princeton, NJ: Educational Testing Service.

Rock, D. 1980. *Disentangling coaching effects and differential growth in the FTC commercial coaching study*. Princeton, NJ: Educational Testing Service.

Schroeder, B. 1992. Problem-solving strategies and the mathematics SAT: A study of enhanced performance. Ph.D. diss., Teacher's College, Columbia University, New York.

Sesnowitz, M., K. Bernhardt, and M. D. Knain. 1982. An analysis of the impact of commercial test preparation on SAT scores. *American Educational Research Journal* 19 (3): 429-41.

Shadish, W., T. Cook, and D. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Shaw, E. 1992. The effects of short-term coaching on the Scholastic Aptitude Test. Ph.D. diss., University of La Verne, La Verne, CA.

Slack, W. V., and D. Porter. 1980. The Scholastic Aptitude Test: A critical appraisal. *Harvard Education Review* 50: 154-75.

Smyth, F. L. 1989. Commercial coaching and SAT scores. *Journal of College Admissions* 123:2-9.

———. 1990. SAT coaching: What really happens and how we are led to expect more. *Journal of College Admissions* 129:7-17.

Snedecor, P. J. 1989. Coaching: Does it pay? revisited. *Journal of College Admissions* 125:15-18.

Stroud, T. W. F. 1980. *Reanalysis of the Federal Trade Commission study of commercial coaching for the SAT*. Princeton, NJ: Educational Testing Service.

Wachter, K. W. 1988. Disturbed by meta-analysis? *Science* 241 (4872): 1407-8.

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

———. 1988. Coaching: Does it pay? Not for Harvard students. *College Board Review* 148 (Summer): 32-35.

Wing, C. W., R. A. Childs, and S. E. Maxwell. 1989. Some field observations of the impact of test preparatory programs on high school students' Scholastic Aptitude Test scores. A report to the Awards Committee for Education and Wake Forest University, Winston-Salem, NC.

Wrinkle, G. W. 1996. A Scholastic Assessment Test preparation class and its effect on Scholastic Assessment Test scores. Ph.D. diss., University of Houston, TX.

Zuman, J. P. 1988. 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, LA, April.

Derek C. Briggs is an assistant professor specializing in quantitative methods and policy analysis in the School of Education at the University of Colorado, Boulder. He teaches courses in statistics and measurement in the Research and Evaluation Methodology program. His research applies statistical and psychometric methods to address a wide range of topics in educational settings. His areas of specialization include causal inference and item response theory. He received his Ph.D. from the University of California, Berkeley, in 2002.