DRAFT OF

24 MAR 10

Touchstones Discussions and Target Practice as Enhancers of SAT Critical Reading Scores.

James F. Guyot

School of Public Affairs, Baruch College CUNY

Training and education are important human capital investments which can be quite expensive. Consequently it is wise to mimic the philosopher who said that the unexamined life is not worth living by advancing the proposition that the unevaluated training program is not worth funding. Recently, such guidance was put forward regarding SAT preparation courses, or coaching, a billion dollar industry in the U.S. focused on raising scores on the SAT without regard to any change in the underlying abilities tested.(Briggs, 2009) This is something akin to target practice that is undertaken with no concern for the ability to bag wild game or to deter aggressors. Evaluations of this sort have long raised interesting methodological problems (Hausknecht, *et al*. 2007). The purpose of this report is to assess the ancillary effects on SAT Critical Reading scores of a novel educational

program, Touchstones Discussions, in a Pre-Collegiate Program in Yangon, Burma, first in a natural experiment and then by comparing the score gains of these Burmese students with a standard of score gain established in the evaluation of SAT preparation courses for high schoolers in the U.S. This undertaking revisits the argument long advanced by the test makers that the better preparation for the SAT is serious study rather than targeted training in the tactics of test taking. (Powers and Camara, 1999)

As a rough initial comparison we could note that for the 780,312 American high schoolers who repeated the SAT one or more times during their junior and senior years, the improvement in verbal scores over a two year span averaged 19.11 points up from an initial average score of 506.99. (Collegeboard, 2004) Correspondingly, for the five waves of 64 Yangon students who took the SAT Reasoning Test between 2005 and 2009, the improvement in Critical Reading scores over the two year span of each wave averaged 40.16 points up from an initial average score of 494.84 . To infer from this difference in outcomes that Touchstones Discussions is a better enhancer of SAT scores than targeted practice with SAT preparation courses, however, would require cautious explanation, which will be undertaken as follows. First, I'll consider several methodological issues by examining three key studies assembled in a recent assertion that SAT

preparation does not accomplish very much. Second, I will present two ways of assessing the additional contribution that Touchstones Discussion makes beyond the background effects of the basso-continuo of test preparation in Yangon, Burma.

COSTS AND BENEFITS OF SAT COACHING

Concerned that the availability of SAT preparation courses (called coaching in the evaluation literature) raised important issues regarding inequality, the National Association for College Admission Counseling commissioned a discussion paper by Dr. Derek C. Briggs which brought in two findings (2009, pp. 12, 19):

1) SAT coaching made a difference of only 5-10 points in the scores earned on the verbal section in subsequent taking of the test, but
2) This small difference could make a significant difference in the chances of admission, particularly to selective colleges.

Such a difference is quite important to the educational rescue mission of the Pre-Collegiate Program in Yangon since this world's longest running military dictatorship is also one of the poorest of the poor countries by UN standards. This means that many quite capable students have no financial ability to attend higher education outside unless provided "full-ride" scholarships of the sort that

are usually available at only the oldest, and often the more selective, four year institutions can provide. Typically, such students have succeeded in an Asian style rote memory educational system where "tuitions" to cram for exams are the rule. Might some more broadly educational pedagogy, say, one that actually develops the ability to think critically about what one reads that the Critical Reading section of the new SAT purports to measure, result in higher scores on that exam than does continuation of the traditional cram school approach?

The coachability of SAT performance has a long history of evaluation studies stretching over half a century (Briggs, 2009) and the College Board has revised the content and format of the SAT twice in recent history in order to reduce its coachability. (Lawrence, et al. 2003) A notable force for the latest, largest reform was the outrage of the president of the University of California, the College Board's largest customer, when he discovered his granddaughter's middle school class drilling on analogies rather than learning substantive content. Consequently, the new SAT introduced in 2005 dropped analogies (and the GRE followed suit this last year).

[Footnote: Contributory factors may have been that the analogies provoked the most complaints from test takers once they got the opportunity to check on items in their scoring and that the analogies subsection was found biased against girls. Notably, the gender gap in average SAT-Verbal score which had been running at

8-9 points in favor of males prior to the reformation dropped to 3-4 points in the new SAT-Critical Reading scores and the absolute male advantage in the number of those scoring above 700 during the previous decade disappeared in the new test.]

METHODOLOGICAL ISSUES IN ASSESSING THE EFFECCTIVENESS OF SAT COACHING

The gold standard for assaying the causal contribution that one event makes toward the appearance of another is, of course, a double-blind experiment with random assignment of cases to treatment and control conditions. The methodological and ethical complexities of such an undertaking in the field of training and education are such that to date no credible attempt of this sort regarding SAT coaching has appeared in publication. We must, then, be content with a wide array of natural or quasi-experiments of varying scientific value. The classic criteria for testing the merit of such attempts were set out by Cook and Campbell (1979, pp. 51-53) as the following "threats to internal validity":

History

If the before to after treatment measurements of the coached group are taken in a different time period from that for the non-coached or less coached group, might some historical change in external conditions enhance or suppress the before or after measurements for one group differently than for the other? Such changes could be a shift in the quality of students admitted in one year

compared with those of another, or a change in motivational tenor from one season to another.

Maturation

If members of the experimental and control groups are at different stages in maturation, a bias could arise from such factors as a natural change in the ability to utilize the benefits of the treatment, as when students become bored or the learning curve flattens for other reasons.

Testing

The experience of taking the test may itself enhance performance in a way that would distort the measurement of pretest and posttest if subjects in one group have had more testing experience than those in the other.

Instrumentation

This is a measurement issue which should not be a problem if the same kind of SAT is used with all groups, although a related measurement issue may arise in determining whether or how much treatment a member of the experimental or control group has been getting.

Statistical Regression

If the self-selected group of those taking SAT coaching is biased toward persons scoring below the mean of those who did not choose coaching, which would be likely since low scorers are the ones who may have a special incentive to try to improve their scores on subsequent testing, then simple regression toward the mean may explain part of any relative improvement shown.

Selection

Other differences between the members of the experimental and control groups may arise since members of these groups were not randomly assigned but rather self-selected. Consequently, several relevant background factors must be assessed in order to test the comparability of the two groups.

THE SAT AND COACHING IN THE U.S.

How do the studies upon which the Briggs report for the NACAC relies stand up to these threats? In an ETS study utilizing the score history of a sample of all high school seniors who took the SAT in the fall of 1995 and all juniors who took it in the spring of 1996, Powers and Rock (1999) conducted a natural experiment. They surveyed those taking the test twice between the two testing

dates in order to ascertain who had taken coaching not offered by their schools. These constituted the experimental group with the remainder taken as controls. Several statistical models were employed in order to approach the rigor of a randomized experiment.

Neither **history** nor **maturation** threatened this undertaking since both the coached and the uncoached took equivalent pre and post tests and took them at roughly the same time.

Testing, however, may have provided a small threat because the coached students were more likely to have taken the PSAT prior to their first experience with an official SAT. The effect of repeated testing generally follows a flattening learning curve, so the coached students' second taking of an official SAT-type test as the pretest in this natural experiment would not have provided them with the same boost in subsequent scores that the first taking in the form of the PSAT had given. By contrast, more members of the control group would have benefitted from the learning experience of their first taking of the official SAT as the pretest in the natural experiment. Such flattening of the learning curve with subsequent repetitions is well established by a meta-analysis of practice effects for tests of cognitive ability in general. (Hausknecht, *et al*., 2007 p. 381) Repeat takers of the

SAT in 2003 and 2004 gained 15 points on the verbal section with their first retake, 11 points on the second, 9 on the third, and 9 on the fourth for the successively smaller groups that took the test an increasing number of times. (Collegeboard 2010) This 15 point boost from a first official SAT experience is probably an understatement of the practice effect since many first time SAT takers had previously taken the PSAT. In a study drawing on the National Education Longitudinal Study of 1988 (discussed below), 68 percent of the SAT takers in 1990-92 had previously taken the PSAT. (Briggs 2001, fig. 1) That 15 point boost in 2003-04 approximates the gain of 12 points that Powers and Rock found for the half a million juniors taking the test in the spring of 1996 who repeated it as seniors in the fall. (Powers & Rock, footnote 3) The effect of this testing threat to internal validity, then, would be to understate the influence of coaching on subsequent scores.

Instrumentation may also be viewed as a threat to internal validity in this study. While the SAT was exactly the same instrument for both coached and uncoached, the questionnaire as an instrument for measuring just how much treatment each group got is less certain. Students were classified as coached or uncoached on the basis of whether or not they "attended coaching programs not offered by their schools." (Powers and Rock, p. 94) Conceding that "test takers …

labeled as 'uncoached' did undertake preparation of various sorts," the authors characterize the coached and uncoached groups as spending vastly different amounts of time in preparation. (footnote 4) By narrowing the treatment difference between the two groups, the effect of this instrumentation threat would be to understate the influence of coaching on subsequent scores.

Statistical regression could be a threat as well since the coached group pretest mean was 6 points lower than the mean for the uncoached. (Table 2) This, too, would have the effect of understating the influence of coaching on subsequent scores. Perhaps this threat was sufficiently controlled by the multiple analytical models employed.

The threat of differential **selection** also was addressed by the six analytical models that the authors assign the task of controlling for "such demographic and background characteristics as sex, ethnicity, parental education, course-taking histories, high school grades, and earlier test scores." (p. 96) There is yet another selection threat to internal validity, one which is unaccounted for in the analytical models that Powers & Rock employ. About one-third of those in the sample frame did not respond. If the coaching process operates differently among those disinclined to respond to the questionnaires, their absence from the sample

analyzed could either overstate or obscure the effect of coaching on test scores. Since nonresponse bias appears to have produced a sample somewhat more able than the general run of SAT takers (Powers & Rock, p. 111) , and since more able test takers are more effective users of any coaching advantage, the likely effect of nonresponse bias would be to overstate the influence of testing on subsequent scores.

Two studies, by Briggs and by Briggs and Domingue, one drawing on the 3,618 members of the National Education Longitudinal Study of 1988 who had taken both the PSAT and the SAT and the other a replication of it drawing on the Educational Longitudinal Study of 2002, follow much the same format as the Powers and Rock study, with several notable differences. (Briggs 2009)

Threats from **history** are still absent yet a new threat from **maturation** arises. In this study the comparison is made of changes between the PSAT and the SAT scores rather than between SAT and SAT. The data are drawn from high school transcripts which report only the highest SAT score. (personal communication with Briggs, 1 March 2010) Since that highest score could come from any one of any number of retakings of the SAT, and since consumers of coaching are generally more ambitious or anxious (Briggs 2001, p. 14) so that they

retake the test sooner and more often, then there could be a maturational difference between coached and uncoached during the time of rapid and varied cognitive development that is the post-puberty years. The net result of less maturation between pretest and posttest for the coached group could be an understatement of the effects of coaching.

The threats from **testing** are rearranged in the later studies. Where Powers and Rock measured pretest and posttest score gains from one SAT to another SAT with some uncertainty about whether the pretest was a first experience, both Briggs studies measured score gain from PSAT to SAT with some uncertainty about whether the coaching treatment occurred in between, or before, or after the pretest and the posttest. (Powers, 2001) Consequently, in both sets of studies uncertainty in the process leads to a likely understatement of the influence of coaching on subsequent SAT scores.

The **instrumentation** threat identified for the Powers and Rock study (varying degrees of treatment among both the experimental and the control groups) was resolved in the Briggs study by analyzing the differences in coaching effect with different degrees of coaching. (Briggs 2001, Table 4) This may bring up a different measurement problem, however, since the data on SAT scores are

from high school transcripts and report only the highest score. While the highest score is the most valid score for predicting college grades (Collegeboard, Sol Arbeiter's reference), the highest reported score of someone who takes the test many times after the PSAT will likely be higher than the maximum score the same individual would obtain on a smaller number of retakings. Since the coached are more likely to take the SAT many times, their maximum score is likely to be more inflated than the maximum score for the uncoached. In this way a bias in measurement may well lead to an overstatement of the influence of coaching on subsequent scores.

By contrast to Powers and Rock, Briggs found his coached students, rather than those who were uncoached, to have higher pretest, i.e., PSAT, scores. (2001, p. 14) This could result by means of **statistical regression** in the understatement of coaching effects.

Whether the threat of differential **selection** engendered by nonresponse bias in the Briggs studies is more or less than in the Powers and Rock study is uncertain since we don't know the nonresponse rate or its biases for the 1990-92 panel of the National Education Longitudinal Survey of 1988. [look for it] We can note, however, that for those in the NELS:88 who took the PSAT, 96.7%

responded to the question of whether and what kind of coaching they received. While the sample size available for maximal controls for selection bias in the linear regression model Briggs employed is reduced by over one-third, an intermediate set of controls reduces sample size only slightly. In this moderately reliable sample the limited set of controls reduces the estimated coaching effect from 14 to 8 points. (2001,Table 5)

Powers and Rock (109) had concluded that "on average coaching seems to effect(sic) SAT I verbal scores by about 8 points." This figure is congenial with the findings in the studies by Briggs. Keeping in mind the caution that most of the threats to internal validity worked in the direction of understating the effect of coaching, I will take something around 10 points on the SAT-CR as a reasonable U.S. standard for comparing with the enhancement of SAT scores attributable to Touchstones Discussions in Yangon.

THE SAT AND TOUCHSTONES DISCUSSIONS IN YANGON

Each April from 2005-09 successive waves of the Pre-Collegiate Program (designated Waves I thru VII), began a ten month program of intensive academic work intended to break students from the Asiatic mode of rote learning at which they had succeeded in a national Matriculation Examination (Matric) following ten years of government schooling or in GCE O-Level examinations prepared for,

in most cases, at the International Language and Business Centre (ILBC), the largest private English language school in the country. The purpose of the Program is to prepare Burmese nationals for modern post-secondary education at liberal arts colleges in the U.S. or at universities in Japan, Canada, Sweden, or Singapore, novel situations in which they would be required to engage in class discussion and write course papers rather than sit in lectures and take written examinations on a set syllabus. The Touchstones Discussions focused on extensive critical reading of a variety of short texts by classes of 10 to 15 students seated at a round table where all were encouraged to participate by a non-directive leader. This mode was central to the Program's pedagogy and suffused the substantive courses as well. The Touchstones Discussions that animated the Pre-Collegiate Program were a truly novel experience during the first few months. They reached a mature stage by the fall and might be expected to have less impact on academic skill development at that stage with a flattened learning curve of the sort noted above with SAT coaching enterprises. I also assume that the cramming with private tutors had reached the phase of a flattened learning curve at this advanced stage in their educational careers since many had been doing it for most of their academic life.

As most students had little family money, they sought full-ride scholarships, which were most available at US colleges old enough to be rich enough to afford such largess. These colleges are usually quite selective in terms of SAT scores. Consequently, from the first waves of the Program on, students took the SAT twice, in October and again in December or January, for purposes of reliability in this crucial admissions criterion. For all but a few it was their first official SAT testing. Fall was a busy time for two tasks -- completing college applications and preparing for the SAT examinations. Many students approached this second task in the traditional fashion that had earned them high scores on the Matric or the GCE -- cramming with private tutors. In order to provide a prediction for individuals' likely SAT scores in the fall, an earlier official SAT testing was inserted at the end of spring with Waves V, VI, and VII. This enlightened the college search process by setting reasonable target schools in terms of SAT selectivity. Thus, the natural operation of the Pre-Collegiate Program provides materials which may be assembled into a natural experiment to test the effect of Touchstones Discussions on SAT Critical Reading scores.

A DESIGN FOR ASSESSING THE EFFECT OF TOUCHSTONES DISCUSSIONS ON SAT SCORES

The question asked in the U.S. studies was whether SAT coaching improves SAT scores and by how much, compared to a steady state of schooling. No analysis was made of variations in the content or rigor of that schooling. I ask a symmetrical question in Yangon, whether Touchstones Discussion improves SAT scores and by how much, compared to a steady state of SAT coaching. Since I assume that coaching of all sorts is so pervasive among ambitious Burmese students that by the time they are in the high teens its effects are fairly far along a flattening learning curve, I make no analysis of variation in the amount or type of coaching. I assess the Touchstones Discussion effect in two ways, first, by means of a quasi-experimental design and then as a simple cross-national comparison.

The quasi-experimental design I employ for assessing the Touchstones Discussions follows a different format that the U.S. studies. In the first instance it is a hybrid design incorporating the Campbell and Cook (1979 pp. 120-123) "Removed-Treatment Design with Pretest and Posttest" as shown schematically in Fig. 1 below:

Fig. 1

Removed-Treatment Design with Pretest and Posttest (modified)

			___	
O_1	X	O_2	X	O_3
SAT- CR	Virgin Soil	SAT-CR	Mature	SAT-CR
May or June	Touchstones Discussions	October	Touchstones Discussions	December or January
			___	
		O_4	X	O_5
		SAT-CR	Mature	SAT-CR
		October	Touchstones Discussions	December or January

The causal inference runs as follows: a greater change between O_1 and O_2, or from pretest to the first posttest, than between O_2 and O_3, or from the first to the second posttest, indicates that the treatment had more effect than did the circumstances where treatment was absent. Cook and Campbell (pp. 122-3) criticize a design similar to this which also used only one observation between the

Treatment and the Removed-Treatment stages because of the possibility of sampling error in the O_2 observation. I address this problem by replicating by means of O_3 and O_4 observations with a different sample that has had both the Treatment and the Removed Treatment but experienced no pretest corresponding to O_1.

FINDINGS IN YANGON

The first observation is a SAT Critical Reading score obtained in May or June, followed by a treatment -- four or five months of critical reading development by means of Touchstones Discussions -- and then a second observation when an SAT is taken in October. I characterize the treatment as a "virgin soil" experience because Burmese students have never been educated in this way before. The Touchstones Discussions continue but reach a more mature stage and do so during a time when the students are focused on preparing college applications. This constitutes the Removed-Treatment condition. The third observation for the Pretest-Posttest group then comes one or two months after the second observation. At this time the Posttest Only group gets their first observation. Let us now put data into this format as in Table 1 below.

Table 1

SAT Scores for Two Groups of Pre-Collegiate Program Students in Yangon

Medians with Means in Parentheses

	Pretest	Score Gain*	Posttest	Score Gain*	Posttest
Pretest-Posttest					
Group	470 (464)	50 (63)	520 (527)	10 (11)	520 (530)
Waves IV, V, VI	n=32		n=32		n=26
Posttest Only					
Group			500 (505)	0 (7)	500 (512)
Waves II, IV			n=31		n=31
Augmented**					

*Score Gain medians are for individual score gains, not for the difference between pretest and posttest medians. The mean score gain from posttest to posttest for the Pretest-Posttest Group is for the 26 students with somewhat lower scores on the first posttest who chose to take a second posttest.
** Four students in Wave VI did not take the pretest and hence were added to the Posttest Only group.

In the Pretest-Posttest group we observe a large virgin soil treatment effect by comparing a median score gain of fifty points from pretest to posttest with the median score gain of only ten points once the treatment has been removed, i.e., it has matured. This treatment effect appears also in the comparison of mean scores (which are less reliable than medians, given the small sample size). This mediocre effect of the matured treatment is confirmed by the replication in the Posttest Only Group. Here the effect is even smaller than in the Posttest Only group, perhaps in part because of regression toward the mean by those in that group electing to take a third testing.

THREATS IN YANGON

History should be no threat to demonstrating a difference between virgin soil and mature treatments since it is a comparison within a single cohort. History might be considered a threat to the validity of the replication of mature treatment effects since the two groups are drawn from two historically earlier waves and three later waves, but the two groups are the same in relevant background characteristics, as noted in the selection section below.

Maturation could be a problem for the comparison of virgin soil with mature treatments since the time intervals between testings are different – four or five months for the virgin soil compared to two three months for the mature condition. As Cook and Campbell note (p. 122) "spontaneous linear changes that take place over a given time period" could bias observed differences between conditions if the intervals are unequal. In order to meet this challenge, I will assume that in the coaching habituated environment of Yangon the rate of change is constant and hence the rate of score gain per months between pretest and posttest and between first and second posttests, provides a valid comparison. During the virgin soil period students gained an average of 13.39 SAT points per month compared to 3.67 points during the mature period.

Statistical Regression is not a threat for the comparisons in the Pretest-Posttest group as no selection was made between pretest and posttest while the selection out of high scorers in the set who took the second posttest would err in the direction of inflating any score gain under the mature treatment condition . An even stronger statistical regression effect is evident further down the line in a third posttest taken by members of both groups. Where the median score for the 63 takers of the second posttest was 510, the seven of them who elected to take

the SAT for a third posttest had scored a median of 470 on the second. They raised their median to 520 on the third posttest.

Testing presents no real challenge as the testing practice effect appears to have atrophied for both the Pretest-Posttest and the Posttest Only groups, as indicated by the replication of the small size of score gains between the first and second posttests.

Instrumentation is no threat since all were taking the same or equivalent SATs and there is no question that they were receiving the same Touchstones Discussions treatment.

However, the standards for **selection** into the program did not change over time and the two groups are essentially the same in such background characteristics as sex, prior educational program, prior academic performance, and TOEFL score. They also had similar college admissions outcomes with the top third in course grades in each wave winning full-ride scholarships to selective U.S. liberal arts colleges, as shown in Table 2.

Table 2

Selective College Admissions

Pretest-Posttest Group (n=32)	Posttest Only Group (n=31)
Bryn Mawr	Davidson (2)
Carleton	Denison
Davidson	Dickinson
Grinnell	Earlham
Kenyon	Earlham
Lehigh	Middlebury
Oberlin	Oberlin
Scripps	Simons Rock
St. Johns, Santa Fe	Whitman
St. Olaf	
Trinity	

A YANGON-U.S. COMPARISON?

The three US studies agree on an SAT Critical Reading score gain that is attributable to coaching of about 8 points. My small study in Yangon suggests that Touchstones Discussions may improve scores by 40 to 50 points over a four

or five month period. This difference in outcomes is quite large, even when making allowance for the threats of understatement of coaching gains that I have noted for the U.S. studies. If we were to use the same metric for calculating score gain in Yangon from pretest to posttest as Briggs employed, the maximum score at any subsequent testing during the high school years, the mean score gain in Yangon rises to 79 points.

What about the best that the US has to offer, the big mules of Kaplan and Princeton Review? Powers & Rock (p. 110) found that one of them, "Company A" or "Company B," more or less demonstrated SAT-CR score gains in the teens but none in SAT-Math, while the other consistently demonstrated score gains in the 30s, but only for SAT-Math.

FURTHER SPECULATIONS

The U.S. studies were of the SAT prior to its latest reformation in 2005, when changes were intended to reduce is coachability. While this difference in instrumentation supports the validity of our pro-Yangon comparison, there are several reasons other than the power of Touchstones Discussion to alter cognitive functioning that may account for the greater enhancement of SAT-CR scores in Yangon. These I consider threats to external validity:

1) Yangon SAT coaching may be even more intense than the best coaching in the U.S. as students in the world's longest running military dictatorship are desperate to get a visa to study abroad.
2) As second language learners of English, Yangon students may constitute an extra virgin soil for SAT Critical Reading enhancement by any means whatsoever.
3) In the U.S. the more ambitious and those from higher SES backgrounds gained more from coaching. The Pre-Collegiate Program students may be socially and academically a more highly selected set within the Yangon environment than is the run of those who take coaching in the U.S.

How else should I interrogate present or potential data from out of Yangon in order to seriously counsel our Program's students and others on the opportunity costs of the SAT coaching habit?

REFERENCES

Briggs, Derek C. (2001) The Effect of Admissions Test Preparation: Evidence from NELS:88. *Chance*, 14, 10-18.

Briggs, Derek C. (2009). *Preparation for College Admission Exams: NACAC Discussion Paper*. Arlington: National Association for College Admission Counseling.

College Board (2010) *Effects of Repeating the SAT: Average Scores for Students Who Took the SAT from One to Five Times During Their Junior and Senior Years.* Downloaded 10 JAN 2010.

Cook, Thomas D. and Donald T. Campbell (1979) *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston: Houghton Mifflin.

Hausknecht, John P., Jane A. Halpert, Nicole T. Di Paolo, and Meghan O. Moriarty Gerrard (2007). *Journal of Applied Psychology*, 92, 373-385.

Powers, Donald E., and Wayne J. Camara (1999). *Coaching and the SAT I: College Board Research Note 06.* NY: College Board.

Lawrence, Ida M., Gretchen W. Rigol, Thomas Van Essen, Carol A. Jackson. (2003) *A Historical Perspective on the Content of the SAT: College Board Research Report No. 2003-3*. NY: College Board

Powers, Donald E. (2001). Comment: Using National Education Longitudinal Study (NELS) Data to Evaluate the Effects of Commercial Test Preparation. *Chance*, 14, 19-21.

Powers, Donald E., and Donald A. Rock (1999). Effects of Coaching on SAT I: Reasoning Test Scores. *Journal of Educational Measurement*, 36, 93-118.