Test Disclosure and Retest Performance on the SAT

Lawrence J. Stricker
Educational Testing Service

The aim of this study was to evaluate the effects of disclosing a Scholastic Aptitude Test (SAT) form on the retest performance of examinees who initially took the disclosed form and subsequently took a different form. Retest performance was compared for three random samples of examinees who took the SAT as high school juniors in the May 1981 administration in New York and then retook it in the October 1981 administration: two experimental groups that were sent the standard set of disclosed material for the May SAT, along with either a noncommittal or an encouraging letter intended to vary their motivation to use the material, and a control group that was not sent anything. The three groups were generally similar in the level and retest reliability of their October scores, indicating that access to the disclosed material had no appreciable effects on retest performance.

Public disclosure of the content of admissions tests, originally mandated by legislation in New York and now a nationwide policy of many admissions testing programs (Brown, 1980; "Test-Takers," 1981), has potentially important consequences for the performance of examinees. Although there has been a great deal of speculation about this subject (see the reviews by Brown, 1980, and Strenio, 1979), data are scarce. It is well established, though, that very few examinees seek the disclosed material for most admissions tests, with the striking exception of the Law School Admission Test (see the review by Linn, 1982).

The only information on the effects of disclosure on test performance comes from a study of the specific recall of disclosed material (Hale, Angelis, & Thibodeau, in press). This experiment, in a classroom setting, found that the examinees achieved substantially elevated scores on special forms of the Test of English as a Foreign Language (Educational Testing Service, 1981b), which consisted of questions already disclosed to the students. These effects occurred regardless of whether the questions were discussed in class, but the extent of the effects depended on the size of the pool of disclosed items: the examinees who had been given many more disclosed questions than what subsequently appeared on the special forms of the test obtained lower scores.

Nothing is known thus far about the broader impact of disclosure in the more realistic situation in which examinees repeat a test after taking an entirely different form of the test, which is subsequently disclosed, and then receiving its questions and their answers. This issue is of considerable practical importance in view of the substantial proportion of examinees who repeat admissions tests (e.g., Donlon & Angoff, 1971) and of the weight that admissions officers attach to retest scores (e.g., Educational Testing Service, 1981a).

In principle, access to the disclosed material in such circumstances—in common with retaking a

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 1, Winter 1984, pp. 81–87

0146-6216/84/010081-07$1.60

test, receiving test coaching, and using test orientation materials, such as guidebooks and practice tests (e.g., Educational Testing Service, 1979, 1980)—has the potential for increasing examinees' familiarity with a test's instructions and content, reducing their anxiety about it, and providing an opportunity for them to drill on specific types of questions (Anastasi, 1981; Messick, 1980). Accordingly, insofar as disclosure has any impact over and above these other influences, subsequent retest scores may be affected in two distinct ways. First, the scores may be elevated. All the possible effects of disclosure that were just mentioned should contribute to score improvement. It is noteworthy that retaking a test and test coaching both produce some score gains (see the reviews by Anastasi, 1981, and Messick, 1980).

Second, the retest scores may not measure the same thing as the initial scores. Greater familiarity with the nature of a test and reduced anxiety should lead to a more veridical assessment of ability, whereas intensive drilling should produce a distorted appraisal (Messick, 1980, 1981). Hence, validity may increase or decrease, depending upon the relative importance of these two kinds of influences. Retest reliability may be lowered in any event, for both influences would reduce the correspondence between initial and retest scores. However, the sparse data that are available on these points, based on initial scores on the Scholastic Aptitude Test (SAT; Donlon & Angoff, 1971) and ordinary retest scores on a different form of this test, suggest that the effects may be relatively small, at least for test familiarization and anxiety reduction. Although these influences should make the two scores diverge, the scores had similar validity in predicting college grades (Olsen & Schrader, 1959); and the retest reliability of the scores is extremely high, approximately .9 (Donlon & Angoff, 1971).

A related matter is that these effects on retest scores, rather than being uniform, may vary systematically with the examinees' characteristics. These include variables (such as unfamiliarity and anxiety) that may lead to poor test performance and be alterable by exposure to the disclosed material, as well as other variables (such as motivation and ability) that may determine access to the material and effective use of it. Data are lacking on this issue.

The primary aim of this study was to evaluate the effects of disclosing a SAT form on the retest performance of examinees who initially took the disclosed form and subsequently took a different form. More specifically, the goal was to determine whether receiving the disclosed test affected the level of retest scores and their retest reliability. A secondary purpose was to explore whether the effects depended on the examinees' characteristics—those that may affect performance and be alterable by exposure to the disclosed material, those that may determine access and use of it, and demographic variables.

Method

Procedure

Three random samples, each consisting of 2,500 examinees, were drawn from those taking the SAT (Form 1Z) in the May 2, 1981, administration in New York. The samples were limited to examinees with the following characteristics, as determined from the registration form and other records about the administration: (1) junior in high school, (2) resident of New York, (3) registered on time for a Saturday administration, and (4) SAT Verbal (V) and Mathematical (M) scores were both available for the administration.

Two of the samples, the Not Encouraged and Encouraged experimental groups, were sent the standard set of disclosed material (the operational items on the test, a copy of the examinee's answer sheet, scoring instructions, and key) that is routinely provided to those who request it. The mailing took place at approximately the same time (June 26 to 30) that the disclosed material was sent to the first of the May examinees who asked for it.

The material for the two experimental groups was accompanied by a letter from the College Board, intended to vary motivation to use the disclosed material. The letters for the two groups differed. The letter for the Not Encouraged group consisted of a single paragraph:

> Although you may not have requested them, I am sending the questions and answers, as well as a copy of your own answer sheet, for those parts on the May SAT that counted toward your scores on the test. The College Board is sending these materials, on an experimental basis, to a cross-section of all students who took the test.

The letter to the Encouraged Group contained the same paragraph plus an additional one:

> In the event that you plan to take the SAT again, you may find these materials useful in preparing for the test. They should help you to become more familiar with the instructions and the kinds of questions used, and may make it possible for you to take the test with greater confidence.

Nothing was mailed to the third sample, the Control Group.

Subsequently, the examinees in each of the three groups who retook the SAT (Form 1Y) in the October 10 to 11, 1981, administration were identified, after excluding three examinees in the Not Encouraged group and four in the Encouraged group to whom the disclosed material could not be delivered. The number of examinees with either SAT-V or SAT-M scores available for the administration were 1,248 for the Control group, 1,229 for the Not Encouraged group, and 1,272 for the Encouraged group. Of these examinees, 87 in the Control group, 59 in the Not Encouraged group, and 62 in the Encouraged group had requested the disclosed material for the May administration.

Measures

SAT scores and background variables were used in the statistical analysis. They were obtained from records for the May and October administrations and from the Student Descriptive Questionnaire completed when the examinee applied to take the SAT at the May administration. The SAT scores were (1) May SAT-V, (2) May SAT-M, (3) October SAT-V, and (4) October SAT-M. The background variables were (1) sex, (2) ethnicity, (3) father's education, (4) mother's education, (5) parent's income, (6) financial need, (7) high school type, (8) high school program, (9) high school rank, (10) high school grade-point average, and (11) educational aspiration.

Statistical Analyses

All statistical analyses were limited to the three samples of examinees who retook the SAT in the October administration and had V or M scores available for the administration. Because of missing data for SAT scores and background variables, the sample sizes fluctuated for the analyses; each analysis was based on all the available data. SAT-V and SAT-M scores were analyzed separately throughout, and parallel analyses were carried out for the May and October SAT scores.

It is important to recognize that although the original samples were comparable by virtue of being drawn randomly, self-selection could have produced differences in the fractions of these samples that were retested—the three samples used in this analysis. Hence, sample differences in the October SAT results may be attributable to differences in the composition of the samples rather than to differences in their retest performance.The influence of this self-selection can be determined by comparisons of the May and October results. Because any sample differences in the May results are presumably due to variations in sample composition produced by self-selection, it can likewise be assumed that similar differences in the October results have the same cause.

Score level. Sample differences in May and October SAT means were assessed by one-way analyses of variance. Differences in October SAT means were also appraised by one-way analyses of covariance, controlling for the pertinent May SAT scores (e.g., May SAT-V was the covariate in the analysis of October SAT-V). Interactions between samples and background variables were evaluated by corresponding two-way (sample by background variable) analyses of variance and analyses of covariance, with a separate analysis for each background variable (dichotomized, where necessary). These two-way analyses were carried out by multiple regression methods, each main effect being adjusted for the other main effect, and the inter-

action being adjusted for all main effects. Interactions between samples and May SAT scores (dichotomized) were also evaluated by corresponding two-way (sample by May SAT score) analyses of variance and analyses of covariance. Interactions with May SAT scores were excluded in the analyses where the same May SAT score was also the dependent variable or covariate (e.g., May SAT-V was excluded in the analysis of variance of May SAT-V and in the analysis of covariance of October SAT-V).

Retest reliability. Sample differences in the product-moment correlations between corresponding SAT scores for May and October were appraised by a χ^2 test (Snedecor & Cochran, 1967). Interactions between samples and background variables were evaluated sequentially by the same χ^2 test:

1. An overall test was made of the correlations in the six subsamples formed by dividing each sample on the basis of a background variable (dichotomized the same way as in the analyses of variance and the analyses of covariance). For instance, in the case of sex, the six subsamples were Male Control, Male Not Encouraged, Male Encouraged, Female Control, Female Not Encouraged, and Female Encouraged.
2. If this test was significant, follow-up tests were made of the correlations in the three subsamples at the same level of the background variable. For sex, one level was Male, and its subsamples were Male Control, Male Not Encouraged, and Male Encouraged; the other level was Female, and its subsamples were Female Control, Female Not Encouraged, and Female Encouraged.

This process was carried out separately for each background variable. Interactions between samples and May SAT scores were evaluated in the same way. Interactions with May SAT scores were excluded in these analyses when the correlations were based on the same May SAT score (e.g., May SAT-V was excluded in the analysis of the correlations of May SAT-V with October SAT-V).

Results and Discussion

Score Level[1]

Analyses of variance of initial scores. The means and standard deviations for the May SAT scores in the three samples and the F ratios for the one-way analyses of variance appear in Table 1. The two F ratios were not significant ($p > .05$); the F ratios for the interactions with the background variables and May SAT scores in the two-way analyses of variance were also not significant. These results indicate that self-selection in the examinees who returned for retesting did not affect the comparability of the samples with regard to their initial performance. This point is reinforced by the analyses of interactions, which established that the similarity of the samples extended to a variety of subsamples.

Analyses of variance of retest scores. The means and standard deviations for the October SAT scores in the three samples, along with the F ratios for the one-way analyses of variance, are also shown in Table 1. Both F ratios were not significant ($p > .05$). Similarly, the F ratios for the interactions with the background variables and May SAT scores in the two-way analyses of variance were not significant.

These consistently negative results strikingly demonstrate that the samples did not differ in their retest scores, even when various subgroups were examined. The present findings, taken together with the uniformly negative results in the analyses of initial scores, imply that access to the disclosed material and the motivation provided by the encouraging letter did not affect the level of retest performance, either for the total samples or the subsamples.

Analyses of covariance of retest scores. The covariance-adjusted means and standard deviations for the October SAT scores in the three samples,

[1]Tables containing the means and standard deviations for the May, October, and covariance-adjusted October SAT scores in the subsamples defined by the background variables and May SAT scores, together with summaries of the corresponding analyses of variance and analyses of covariance, are available from the author.

Table 1
Means and Standard Deviations for Initial, Retest and Covariance-Adjusted Retest Scores

	SAT-V			SAT-M		
Sample	Initial	Retest	Adjusted Retest	Initial	Retest	Adjusted Retest
Control						
N	1248	1209	1209	1248	1203	1203
Mean	449.05	465.43	466.72	501.05	506.51	506.19
SD	97.69	101.06	49.65	100.89	104.86	51.55
Not Encouraged						
N	1229	1194	1194	1229	1190	1190
Mean	451.97	470.55	468.74	501.14	509.34	507.60
SD	93.45	97.13	48.07	93.91	100.17	53.69
Encouraged						
N	1272	1225	1225	1272	1229	1229
Mean	450.55	465.73	466.22	497.96	504.78	506.40
SD	94.75	97.82	47.30	99.42	101.76	52.89
F Ratio	.29	1.02	.92	.43	.61	.24

Note. None of the F ratios are significant ($p > .05$).

as well as the F ratios for the one-way analyses of covariance, are also reported in Table 1. These F ratios were not significant ($p > .05$). In addition, the F ratios for the interactions with the background variables and May SAT scores in the two-way analyses of covariance were not significant.

These findings are congruent with the preceding results for the October SAT scores in demonstrating that the samples and subsamples did not differ in their retest scores and in suggesting that the disclosed material and the motivating letter uniformly failed to have an impact on the level of retest performance. The close resemblance between the two sets of results is not surprising, even though the present analyses take into account initial differences in the samples and the other analyses do not, for the samples were observed to be similar in the analyses of May SAT scores.

Retest Reliability[2]

The correlations between the May and October SAT-V scores in the three samples appear in Table 2 together with the χ^2's. The χ^2 was significant ($p < .05$) for SAT-M but not for SAT-V.

In the analyses of the SAT-V correlations in the subsamples defined by the background variables and May SAT scores, none of the χ^2's for the correlations in the subsamples at the same level of these variables was significant. In the parallel analyses of the SAT-M correlations, the χ^2's were significant for one level of Sex (Male), Ethnicity (Nonwhite), and High School Rank (Top Fifth of Class). The statistics for these subsamples are also reported in Table 2.

These results indicate that the sample differences in retest reliability were very minor, being limited to extremely small divergences for SAT-M. The subsample findings also suggest that these differences were not uniform throughout the samples but stemmed from a few isolated subgroups of examinees. Whether this outcome is traceable to variations in sample composition produced by self-selection or in retest performance cannot be determined. In any event, it appears that the disclosed material and the letters had no more than a negligible impact on retest reliability for the samples as a whole as well as for the various subgroups.

[2]Tables containing the correlations between corresponding May and October SAT scores in the subsamples defined by the background variables and May SAT scores, together with the corresponding χ^2's, are available from the author.

Table 2
Correlations between Initial and Retest Scores for Samples and Selected Subsamples

Score and Sample or Subsample	Control N	Control r	Not Encouraged N	Not Encouraged r	Encouraged N	Encouraged r	χ^2
SAT-V							
Total Sample	1209	.87	1194	.87	1225	.88	.46
SAT-M							
Total Sample	1203	.87	1190	.84	1229	.85	6.35*
Subsample:							
Male	609	.88	576	.84	592	.86	6.29*
Nonwhite	127	.90	127	.84	129	.91	6.44*
Top Fifth of Class	438	.85	427	.80	408	.85	7.62*

Note. All the correlations are significant ($p < .01$).
*$p < .05$

Conclusions

The main conclusion of this study is that access to the disclosed test material had no appreciable effects on the subsequent retest performance of examinees in general and various subgroups of them, regardless of whether the performance was defined in terms of the level or retest reliability of the new scores. It also appears, though the evidence on this point is less direct, that use of the material had no discernible effects either. These outcomes are especially noteworthy in view of the large samples involved and the accordingly high level of statistical power that they provide to detect even very small effects.

Although this investigation was concerned with both access to the disclosed material and its use, the data on the latter issue were indirect because of the nature of the experimental design. Access was guaranteed in both experimental groups; use was not ensured in either of them, though an effort was made to increase use in the Encouraged group. Hence, the failure to find effects for this group implies that use of the material had no impact, but does not demonstrate it, in the absence of information about actual use by the examinees.

The general failure to find any effects in this study casts some doubt on the line of reasoning that suggested retest performance might be altered. This reasoning rests on two propositions: (1) greater familiarity with the nature of a test, reduced test anxiety, and intensive drilling on the test's items enhance performance; and (2) these factors are influenced by using the disclosed material. It is entirely possible that the second proposition is incorrect, at least in the present context, because of the nature of the examinees and the test involved. First, the examinees may have been maximally familiar with the test and minimally anxious about it by the time that they received the disclosed material. All had taken the SAT in the May administration and routinely received *Taking the SAT* (Educational Testing Service, 1979), an orientation booklet that contained a sample form of the SAT, when they registered for that administration. Hence, the familiarization and anxiety reduction provided by using the disclosed material may have already been accomplished. This speculation is consistent with the finding that practice on a test and other kinds of exposure to it have the greatest effects on score level for naive examinees and that the gains diminish with repeated practice and exposure (see the reviews by Bond, 1981, and Messick & Jungeblut, 1981). Second, the SAT may not be influenced by drilling because this test includes few, if any, of the kinds of items on which performance can be improved by such practice. Systematic efforts are made to eliminate such items from the SAT (Donlon & Angoff, 1971).

The negative results necessarily raise questions about the efficacy of the experimental operations, particularly the encouraging letter and the variables

used to form the subgroups. The effectiveness of the letter in increasing use of the disclosed material is suggested by the impact that the model for this letter had in a study of test familiarization involving the Graduate Record Examinations Aptitude Test (Conrad, Trismen, & Miller, 1957): the original markedly affected both the amount of time that the examinees used the familiarization material and their test scores (Powers & Swinton, in press). The measures used to define the subgroups in the present investigation, though adequate for exploratory purposes, are not ideal. The May SAT scores are reasonable indexes of the ability to take advantage of the disclosed material. The background variables are adequate measures of key demographic characteristics, but no more than substitutes for direct assessments of anxiety, motivation, familiarity with tests, and so forth.

References

Anastasi, A. (1981). Diverse effects of training on tests of academic intelligence. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 5–19). San Francisco: Jossey-Bass.

Bond, L. (1981). Bias in mental tests. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 55–77). San Francisco: Jossey-Bass.

Brown, R. (1980). *Searching for the truth about "truth in testing" legislation*. Denver CO: Education Commission of the States.

Conrad, L., Trismen, D., & Miller, R. (Eds.). (1977). *Graduate Record Examinations technical manual*. Princeton NJ: Educational Testing Service.

Donlon, T. F., & Angoff, W. H. (1971). The Scholastic Aptitude Test. In W. H. Angoff (Ed.), *The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests* (pp. 15–47). New York: College Entrance Examination Board.

Educational Testing Service. (1979). *Taking the SAT*. New York: College Entrance Examination Board.

Educational Testing Service. (1980). *4 SATs*. New York: College Entrance Examination Board.

Educational Testing Service. (1981a). *ATP guide for high schools and colleges 1981–82*. New York: College Entrance Examination Board.

Educational Testing Service. (1981b). *TOEFL test and score manual, 1981 edition*. Princeton NJ: Educational Testing Service.

Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (in press). Effects of test disclosure on performance in the Test of English as a Foreign Language. *Language Learning*.

Linn, R. L. (1982). Admissions testing on trial. *American Psychologist, 37,* 279–291.

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.

Messick, S. (1981). The controversy over coaching: Issues of effectiveness and equity. In B. F. Green (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 21–53). San Francisco: Jossey-Bass.

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

Olsen, M., & Schrader, W. B. (1959). *The use of preliminary and final Scholastic Aptitude Test scores in predicting college grades* (ETS SR 59–19). Princeton NJ: Educational Testing Service.

Powers, D. E., & Swinton, S. S. (in press). Effects of self-study for coachable test items. *Journal of Educational Psychology*.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames IA: Iowa State University.

Strenio, A., Jr. (1979). *The debate over open versus secure testing: A critical review* (National Consortium on Testing Staff Circular No. 6). Cambridge MA: Huron Institute.

Test-takers may ask for and get answers to SAT's next year, College Board decides. (1981, April 6). *Chronicle of Higher Education,* pp. 1, 10.

Acknowledgments

This study was supported by the College Entrance Examination Board. Thanks are due Donald L. Alderman, John A. Centra, Philip K. Oltman, Donald E. Powers, Donald A. Rock, and Warren W. Willingham for advising about the research design and statistical analysis; Donald Schiariti for supervising the assembling and mailing of the disclosed material; Peter E. Smith for arranging the retrieval of the data; Patricia W. Cox for statistical calculating; Norma A. Norris for computer programming; and Gordon A. Hale, Donald E. Powers, and Gretchen W. Rigol for critically reviewing a draft of this article.

Author's Address

Send requests for reprints or further information to Lawrence J. Stricker, Educational Testing Service, Princeton NJ 08541, U.S.A.