DOCUMENT RESUME

ED 382 651 TM 023 085

AUTHOR	Marco, Gary L.; And Others
TITLE	Trends in SAT Content and Statistical Characteristics and Their Relationship to SAT Predictive Validity.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-90-12
PUB DATE	Aug 90
NOTE	276p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	Change; *College Entrance Examinations; Correlation; *Educational Trends; Grades (Scholastic); Higher Education; *Predictive Validity; *Scores; Statistical Distributions; Test Content; Test Format; *Test Results; Trend Analysis
IDENTIFIERS	College Entrance Examination Board; *Scholastic Aptitude Test

ABSTRACT

Data from the College Board Validity Study Service show that the average multiple correlation of the Scholastic Aptitude Test (SAT) with college grades peaked in 1974 and then tended to decline. Data from other sources also estimate a small average decline from 1974 to 1985. This study documented changes in the SAT and related these changes to trends in the predictive validity of the SAT, focusing on changes in test format, content specifications, statistical characteristics of the test, and equating procedures associated with SAT test forms taken by classes graduating from high school in 1971 through 1985. Data came from November and December test forms for these years. Analyses indicated that the slight decline in SAT predictive validity was not due simply to the shortening of the SAT, nor to changes in content or statistical characteristics, nor to changes in equating methods. No patterns of change in the various indices that were reviewed were consistent with the patterns in validity. Thirty-nine tables in the text, 17 in an appendix, and 14 figures illustrate these analyses. Appendixes present SAT directions and sample questions and contain supplemental tables. (Contains 33 references.) (SLD)

RR-90-12

RESEARCH REPORT

TRENDS IN SAT CONTENT AND STATISTICAL CHARACTERISTICS AND THEIR RELATIONSHIP TO SAT PREDICTIVE VALIDITY

Gary L. Marco
Carolyn R. Crone
James S. Braswell
W. Edward Curley
Nancy K. Wright

ETS

Educational Testing Service
Princeton, New Jersey
August 1990

Trends in SAT Content and Statistical Characteristics and Their Relationship to SAT Predictive Validity

Gary L. Marco
Carolyn R. Crone
James S. Braswell
W. Edward Curley
Nancy K. Wright

Educational Testing Service
Princeton, New Jersey

CONTENTS

TABLES

FIGURES

ACKNOWLEDGMENTS

We are grateful to Warren Willingham for his encouragement and guidance throughout the preparation of this report. Additionally, we thank Nancy Burton, Linda Cook, Ernest Kimmel, and Cathy Wendler for their careful and thoughtful reviews, often under unrealistic time constraints, at various stages of report preparation.

We are indebted to Gaye Allen for her perseverance and dedication to this project. Gaye typed the tables and text of this report and patiently responded to an untold number of requests for revisions. We also thank Susan Bryce for the many hours she spent producing the figures for this document; Mildred Rowen for classifying the items in 30 SAT forms by content characteristics; and Margarete Zicha for entering thousands of data points into the computer data base and creating dozens of tables. We also appreciate the efforts of Elizabeth Carrick, who edited the report and suggested many changes that improved the text, tables, and figures.

The following persons also with helped the project: Carol Jackson, who helped organize the work of test-development staff on the project and reviewed several SAT forms; Ted Blew and Karen Carroll, who established the basic structure of our data base; David Wright, who drafted a section of the equating part of the report; Sydell Carlton, Pamela Cruise, Jeff Haberstroh, Patricia Klag, Mel Kubota, Geraldine May, Marilyn Sudlow, and Beverly Whittington, who reviewed several tests and provided impressions of the comparability of these forms over time; and Jean MacFarland who provided administrative and secretarial assistance for test-development staff.

Finally, we thank Linda Johnson and Kathy Howell for providing administrative support, and Samuel Messick and his staff for their efforts in organizing the review process.

1. OVERVIEW

Data from the College Board Validity Study Service (VSS) show that the average multiple correlation of the Scholastic Aptitude Test (SAT) with college grades peaked in 1974 and then tended to decline. These data, however, came from colleges that participated voluntarily in the VSS and thus are difficult to interpret. Recent estimates of trends in SAT predictive validity are available from Ramist and Weiss (in press), who based their estimates on comparable groups of students and colleges. They estimated the average decline to be about .04 to .05 from 1974 to 1985, about the same as the average increase from 1970 to 1974.

Even though the decline in predictive validity coefficients was not large, it is important to investigate the extent to which validity trends might have reflected changes in the test itself rather than other factors (e.g., changes in courses taken by college students). The purposes of this study were (a) to document changes related to the SAT and (b) to relate these changes to trends in SAT predictive validity. This study focused on changes in (a) test format, test content specifications, and actual test content, (b) statistical characteristics, and (c) equating procedures associated with SAT test forms taken by classes graduating from high school in years 1971 to 1985.

The study covered the period from 1971 to 1974 as well as the 1975-85 period to represent the times when average correlations with college grades increased and decreased. The SAT item and test data used in the study came from forms administered in November and December of the senior year to members of the high school graduating cohort from 1971 to 1985. These forms accounted for over half of the scores of college-bound seniors who graduated from high school in this period.

Changes in Test Content and Format and in Test-Development Procedures

Primary among the changes that occurred in the test during the period studied was the shortening of the SAT-Verbal (SAT-V) and the SAT-Mathematical (SAT-M) in October 1974 from 75 to 60 minutes each. From 1970 to 1974 the SAT-V and the SAT-M each consisted of one 30- and one 45-minute section and were administered in one of two fixed section orders at each administration. In October 1974 the SAT-V and SAT-M were shortened to two 30-minute sections each to permit the introduction of the Test of Standard Written English (TSWE) into the testing program. The revised SAT-V consisted of 85 items rather than 90 items; the SAT-M, like its predecessor, consisted of

60 items. To maintain high test reliability, test developers increased the proportion of items that could be answered at a faster rate by adding more Antonyms in the SAT-V and replacing Data Sufficiency items with Quantitative Comparisons in the SAT-M.

In addition, test developers reduced the number of Reading Comprehension items in the shortened SAT from 35 to 25 by eliminating the Synthesis passage and one of the two science passages--leaving five reading passages in the shortened SAT. In October 1978, when three shorter reading passages replaced two longer passages, the second science passage was returned to the test.

The order in which the sections of the SAT and the TSWE were administered changed as well as the numbers and types of items in the shortened SAT. The six SAT and TSWE sections, including the "variable" section used for pretesting, were in most instances administered in six fixed section orders at each administration from 1974 to 1978. Then, beginning in October 1978, the SAT was administered in one of two fixed section orders at each administration. Finally, from the 1980-81 testing year to the 1984-85 testing year, one of three section orders was used at each administration.

Increased sensitivity to the interests of minority groups and females resulted in other changes. One minority-relevant reading passage was included in at least one form of the SAT-V from testing year 1972-73 to testing year 1976-77. From December 1977 on, each new form of the SAT-V has included a minority-relevant passage. Moreover, the generic "he" was virtually eliminated from the SAT-V by the 1977-78 testing year. Formal test sensitivity review began in 1980 to ensure that test forms were free of material offensive or patronizing to females and minority groups.

The only other change of note resulted from the passage of legislation in the State of New York, which required the disclosure of five SAT forms per calendar year beginning with 1980. As a result more new SAT forms had to be produced. Test developers constructed seven rather than five new SAT forms in testing year 1979-80, and ten in testing year 1981-82. The number of test-development staff members increased correspondingly to cope with the new work load.

Such changes in test content and format could conceivably have affected the predictive validity of the SAT, particularly if the various item types used in the SAT were differentially valid. However, special studies (Schrader, 1973; Schrader, 1984; Burton et al., 1989) indicated that no one item type on the SAT-V or the SAT-M predicted college grades with any more accuracy than the other item types. Therefore, changes in content probably did not affect predictive validity.

Changes in Statistical Characteristics

Statistical specifications were changed twice in the 1971-85 period. With the introduction of the shortened SAT in October 1974, test difficulty was reduced somewhat. The proportion of difficult items was either maintained (for the SAT-M) or increased slightly (for the SAT-V), while the proportion of easy items was increased somewhat in an attempt to lower difficulty and still maintain measurement power at the upper end of the score range. In terms of pretest statistics, the mean item-total biserial correlation was increased from .42 to .43 for the SAT-V and remained at .47 for the SAT-M. In January 1982 the statistical specifications for the SAT-V were again revised by reducing the proportion of difficult SAT-V items and increasing the proportion of moderately difficult items.

Analyses of the data from the November and December forms showed that actual mean item difficulty deviated slightly from specified values. SAT-V forms tended to be harder than specified prior to 1974 and easier than specified from 1974 on. SAT-M forms, on the other hand, tended to be easier than specified prior to 1974, and sometimes easier and sometimes harder from 1974 on. Although there was some form-to-form variation, the average item-difficulty distribution corresponded closely to specifications. In addition, both SAT-V and SAT-M mean item-total biserial correlations tended to be higher than specified, particularly those for the SAT-M.

Another measure of test difficulty, the score conversions that result for score equating, were somewhat inconsistent with the actual mean equated deltas. The score conversions indicated that the SAT-V and the SAT-M forms given in the more recent years of the period tended to be easier than previous forms.

The relative difficulty of the test for the test takers was measured by the mean adjusted proportion correct (raw score mean divided by the number of test items) and by the mean observed delta (the standardized measure of item difficulty used at ETS). Both indices showed that the November and December SAT forms were difficult for the average test taker. The November test SAT-V and SAT-M forms tended to become easier for the test takers over time. This trend would tend to improve the measurement power of the test for the middle-to-low-scoring test takers and presumably for the average student in a validity-study sample.

Measures of test speededness showed that the longer of the two SAT-V sections tended to become gradually more speeded from 1974 on for November test takers and from 1973 on for December test takers. The other SAT-V section was relatively unspeeded except in 1974. The speededness of the two SAT-M sections tended to decrease or remain stable as time went on. Decreasing speededness would tend to improve the measurement power for middle-to-low- scoring test takers, and possibly reduce measurement power for high scorers. Predictive validity might improve as well for the typical college conducting validity studies. Because the changes in speededness were not large, it is unlikely that changes in speededness caused any change in validity coefficients during the 1971-85 period.

Reliability coefficients tended to be very high--at or above .91--for both the SAT-V and SAT-M from 1970 to 1984. Although some reliabilities dropped in 1974, reliabilities from 1974 on tended to be at least as high as those during the 1970-73 period. Test-retest correlations for spring test takers who repeated the test were also high, ranging from .87 to .89. No trends were apparent. Lower reliability coefficients will attenuate validity coefficients, but the slight changes that occurred in reliability of the November and December forms were too small to affect predictive validities in the .30 to .40 range.

Correlational patterns among the SAT-V, the SAT-M, and the TSWE varied somewhat from year to year. Still, a slight downward trend appeared in the correlations between the SAT-V and the SAT-M, which decreased from .68 to .66 for both November and December test takers during the 1970-84 period. Most of the correlations of the SAT-V with the TSWE fell between .78 and .79. Those for the SAT-M with the TSWE ranged between .62 and .64. Over time both sets of correlations increased slightly and then decreased to previous levels. Correlations between the two SAT-V sections and between the two SAT-M sections ranged from .97 to 1.00 after correction for attenuation. The corrected correlations for the SAT-M tended to be slightly higher than those for the SAT-V, indicating somewhat greater homogeneity among SAT-M item types. No pattern was evident for the SAT-V, while the SAT-M became slightly more homogeneous from 1974 on. Corrected correlations between the Reading and Vocabulary subscores were also very high--ranging from .92 to .96. The correlations appeared to be more stable after 1978. The relative stability in these correlations indicates that the changes in the test had little effect on correlational patterns. Certainly, the few trends observed in the data would have only a minor effect, if any at all, on predictive validity.

Changes in Equating Procedures

Equating is the process by which scores from different forms of the SAT are placed on scale. Throughout the period studied, the anchor-test equating design was in use. Under this design, each form of the SAT-V or the SAT-M is equated back to two old forms through common items, usually administered together in the variable section in the SAT test booklet.

The same linear equating methods (Tucker and Levine) were used from 1970 to January 1982. In January 1982, because of the change in statistical specifications for the SAT-V, item-response-theory (IRT) equating began to be used, sometimes in combination with linear methods, for both the SAT-V and the SAT-M.

For any given equating, two individual equating lines are averaged to produce an operational equating line. From 1970 to 1981, all of the November SAT-V and SAT-M operational equating lines came from an average of two Tucker equating lines. On the other hand, many of the December operational equating lines came from an average of Tucker and Levine lines or Levine and Levine lines. (Levine equating is used whenever large ability differences are observed between new- and old-form equating samples.) There was no consistent trend in the use of the Levine equating method from 1970 to 1981.

From 1982 on two IRT (curvilinear) equating lines were usually averaged to produce the operational score conversions, but in some cases a linear line was averaged with an IRT line. The introduction of IRT equating into the testing program does raise the question as to what effect the adoption of a new equating method might have had on validity. This appropriateness of equating methods was addressed in special analyses, and the choice of equating method was shown not to have much of an effect on test scores. Various indices related to equating were reviewed for November and December SAT forms administered from 1970 to 1984. These indices took account of differences between the abilities of the new- and old-form samples, between equating-test-total-test correlations, and between the two individual equating lines from which a given operational equating line is derived. In addition, a composite equating index--a weighted combination of the other indices--was also calculated for the equatings. Although there were some outlying values and at times indices fluctuated more than at other times, no notable trends occurred from 1970 to 1984 in any of the individual indices. The November and December patterns for the SAT-V composite index were variable and inconsistent. Although the November and December composite indices for the SAT-M tended to be relatively low in the 1981-84 period, lower values were observed in December from 1973 to 1975. To summarize, no deteriorating patterns were observed in the various equating indices. Thus, it is not likely that changes in equating procedures affected the predictive validity of the SAT.

Two special analyses were conducted to determine whether it was appropriate to use linear equating in 1974 and IRT equating in January 1982--times when statistical specifications changed. Operational and equipercentile equating lines were compared at the raw score midpoint for November and December forms administered from 1970 to 1984. The operational and equipercentile scaled scores at the raw-score midpoints deviated only 2 to 3 points from one another The small sizes of the deviations suggest that the linear equating was appropriate in most cases from 1970 to 1981. Differences between linear and curvilinear equating lines at the midpoints for forms equated in the 1970-81 period were similar to differences between two curvilinear lines for forms equated from 1982 to 1984. Thus, if scores had been based on equipercentile (curvilinear) equating, they would not have differed much from the scores based on operational equating methods.

Linear and curvilinear score conversions for the full score range were compared for November and December 1974 forms and for the two January 1982 forms. These comparisons show that two sets of scores correlated at least .998 with one another. The differences were not large in the middle part of the score scale, and the means and standard deviations of reported scores would not have changed much had the alternative equating method been used.

These results indicate that linear or curvilinear equating would have yielded similar reported scores, particularly in the middle part oi the score scale. Apparently, the choice of equating method was not a critical determinant of scores. If the scores themselves would not have changed much, had an alternative equating method been used, then neither would their relationships with other variables like freshman grades.

Perhaps the most important evidence of the integrity of the equating process came from SAT scale stability studies. Modu and Stern's studies (1975, 1977) of scale stability from 1963 to 1973 showed that the scales for SAT-V and SAT-M drifted upward by about one and a half points a year. McHale and Ninneman (1990) reported that from 1973 to 1984 the SAT-V scale remained stable. Their study yielded mixed results for SAT-M, but the scale drift at worst is estimated at no more than one and a half points a year. In any case, the shifts were small and point to the integrity of the score-equating process.

Conclusion

The key aspect of this study, besides documenting noteworthy changes in the SAT over time, was to see to what extent trends in validity are associated with changes in the test. The analyses of item and test statistics from November and December SAT test administrations from 1970 to 1984 indicated that the slight decline in SAT predictive validity was not due simply to the shortening of the SAT, nor to changes in content or statistical characteristics, nor to changes in equating methods. No patterns of change in the various indices that were reviewed were consistent with the patterns in validity.

2. BACKGROUND OF THE STUDY

The College Board has monitored trends in the predictive validity of the SAT and high school record since 1964, when the VSS began to conduct validity studies for colleges. The VSS performs data analyses relating SAT scores and high school record (class ranks or grades) to college freshman grades provided by participating colleges. SAT predictive-validity data from the VSS show considerable variation over time.

Because participation in the VSS is entirely voluntary, any trends in the data could be due to the different types of colleges that participated in the VSS from year to year. College participation rose particularly in 1972 when the VSS began to use score data and student-reported high school record from testing-program files. Changes in the ranges of scores within institutions over time is another factor that could cause misleading trends.

First Morgan (1989) and then Ramist and Weiss (in press) examined trends in average validity coefficients when the results are based on comparable groups of students and colleges. The adjusted data show less pronounced trends than do the unadjusted, but all data show the same general pattern--the correlations between SAT scores and college freshman grades went up early in the 1970's and then declined by roughly the same amount from 1974 to 1985.

Table 1 reports estimates of the average correlations of the SAT with college freshman grade point average from Ramist and Weiss (in press) for classes entering college from 1970 through 1985. (The tables and figures are located at the end of the report to permit ready reference, as many of them are referred to several times in the report.) These correlations are more accurate than the unadjusted correlations based on colleges that chose to participate in the VSS in any one year because they are based on comparable samples of students and colleges. The estimated correlations were derived from paired data on colleges that conducted more than one validity study through the VSS from 1970 to 1985 and are adjusted for selectivity in college admissions and enrollments. The table shows that the average adjusted multiple correlation of SAT-V and SAT-M scores with college freshman grades increased in an uneven fashion in the early 1970's and then gradually decreased from 1974 to 1985.

Despite the fact that the real validity decline from 1974 to 1985 appears to be about half of what the average unadjusted multiple correlations indicate, it is still important to examine the extent to which changes that occurred in the SAT might have affected predictive validity. To what extent did the validity trends reflect changes in the test itself rather than changes in other factors (e.g., changes in the courses taken during the freshman grades)? The purposes of this study are (1) to document changes in the test, the test's statistical characteristics, and equating procedures for SAT-V and SAT-M test forms taken by classes entering college from 1971 to 1985 and (2) to relate these changes to trends in SAT predictive validity. In particular, this study addresses the following questions about changes in the content and statistical characteristics of SAT forms:

1. What changes occurred in test content, test format, and test-development procedures?
2. What changes occurred in the statistical specifications and in the statistical characteristics of the test forms?
3. What changes occurred in equating methods?
4. How might any such changes be related to trends in SAT predictive validity?

The study covers the period from 1971 to 1974 as well as the 1975-85 to represent the period during which validity increased as well as the period of the decline. The study stops with the year 1985 because that was the last year for which VSS validity data were available on a reasonably large sample of colleges. Relatively few colleges conducted studies through the VSS for classes entering college in 1986 and 1987, and validity data for the entering class of 1988 are only now becoming available. The decrease in validity studies for 1986 and 1987 entering classes was due to an overhaul of the VSS computer system, which meant that on a temporary basis participating colleges had to provide SAT scores and high school record from their own files rather than obtain them conveniently through the VSS.

The test forms most relevant to this study were those taken by the classes entering college from 1971 to 1985, the classes that provided data for validity studies. These classes consisted primarily of seniors who graduated the previous spring. Most of their SAT test scores came from forms administered between March of the junior year and January of the senior year. This study used data provided by the forms administered in November and December. The November administration alone provided approximately 30% to 40% of the scores for the graduating cohort; the November and December administrations together provided over half of the scores (see Tables 2 and 3).

Tables 2 and 3 give the numbers of test takers, means, and standard deviations for high-school seniors who took the SAT in November and December from 1970 to 1984. They also report the same data for the full senior cohort of SAT takers in the graduating classes the following spring. For comparison purposes the tables provide the means and standard deviations for the entering freshman classes from colleges that participated in the VSS. The freshmen represented in validity studies scored higher on average than the seniors who took the November and December SAT forms. The November senior means were higher than the December means and in general more like the cohort means. The average of the means for the two administrations, however, was closer to the mean for the entire cohort than was the mean for either of the individual administrations.

The test data for the study came from test-analysis samples and equating samples for the November and December forms administered from 1970 to 1984. Later sections of this report refer to item and test statistics on these samples. No attempt was made in this study to estimate the standard errors of the statistics, partly because of the complexity involved, but mainly because they would in all likelihood be small given that statistics came from large samples. Therefore, any differences observed over time are interpreted as real differences. The essential question for the study is not whether the observed variation is statistically significant (i.e., whether it is larger than the variation one would expect from random sampling). Rather, it is whether the observed variation shows any trends that are practically significance and are related to validity.

This study focuses on changes in test-related variables and not on changes in performance of the group that happened to be taking the test. For some variables used in the study (e.g., speededness and reliability indices), changes in the test were inseparable from changes in the characteristics of the test takers. The study addresses such changes, even though in such cases it is not clear whether test-related or group-related factors caused particular indices to change.

The approach taken in this study is to describe changes related to the test and their possible effect on validity in three parts: (1) changes in content, format, and test-development procedures, (2) changes in item and test statistics, and (3) changes in equating methods. This approach is awkward at times because the changes were not independent, and often occurred at the same time. For example, content and statistical specifications both changed in October 1974 when the shortened SAT was introduced. As a result, the reader will encounter some overlap in the discussions. The discussion of the relationship of changes associated with the test to trends in SAT predictive validity comes at the end of each of the three parts.

In relating changes in the test to trends in validity, the basic strategy followed here is to look for similar patterns of change. Many of the changes in the test occurred all at once and therefore would be expected to produce a single, sharp increase or decrease--both in the test's statistical characteristics, such as reliability, and in predictive validity. Presumably, if inconsistent patterns were found, then changes in the test could not have accounted for changes in validity.

Two outcomes would lead one to conclude that changes in the test had little to do with trends in predictive validity: (1) The test changed when predictive validity remained stable; (2) predictive validity changed when the test remained stable. Since the 1971-85 period was a time of continually changing validity, the question is, Are there times during this period when the test itself did not change? If so, changes in the test, by themselves, could not have caused the changes in predictive validity.

This study concentrates on whether major changes associated with the test could have affected predictive validity. It is possible that small, subtle changes in the test, spread out over a number of years, might have produced a gradual change in the statistical characteristics of the test and in predictive validity. Other research in progress (Angoff, Pomplun, McHale, & Morgan, in press) is addressing this possibility by comparing the predictive validities of old and new test forms on the same samples of students.

3. CHANGES IN TEST CONTENT AND FORMAT AND IN TEST-DEVELOPMENT PROCEDURES

This section of the report documents changes that have occurred to the SAT in terms of test content, test format, and test-development procedures from 1970 to 1985. This section also includes a brief discussion of changes in test familiarization material. In addition, it includes a review of selected test forms from the 1970-85 period. Test developers documented changes that they saw among the forms, being particularly alert to changes that were not part of the content and statistical specifications for the test.

Test-development files, including actual test copy, provided most of the information on changes in the test content and format. The technical manuals for the Admissions Testing Program (Angoff, 1971; Donlon, 1984) provided some additional information about such changes. Other source material included student booklets and the recent report by Cruise and Kimmel (1989), which provided information about trends in content and gender references in the

SAT-V from 1961 to 1987.

Changes in Content

SAT-Verbal

There have been two revisions to SAT-V content specifications during the period covered by this study (see Figure 1). The first occurred in 1974 when the SAT was shortened to allow for the introduction of the Test of Standard Written English (TSWE). The second, a less significant revision than the first, resulted from the decision in 1978 to use some shorter reading passages on the verbal test. A summary of the three sets of specifications appears in Table 4.

From 1961 to 1974 the SAT-V was a 75-minute, 90-item test. The test was administered in one 45-minute section and one 30-minute section. Content specifications called for each form of the test to contain approximately 60% discrete items (Sentence Completions, Antonyms, and Analogies) and approximately 40% items based on reading passages. Directions for each item type with examples are given in Appendix A. Seven reading passages were included, one from each of the "ollowing classifications: Narrative, Biological Science, Physical Science, Argumentative, Humanities, Synthesis, and Social Studies. (The Synthesis passage dealt with the relationship between the sciences and the humanities.) Neither a minimum nor a maximum length was specified for any particular passage, but the seven passages together were not to contain more than 3,500 words.

Beginning with the October 1974 administration, the SAT-V was shortened from 75 to 60 minutes and from 90 to 85 items. The new test was administered in two 30-minute sections. The proportion of the test devoted to the most time-consuming item type, Reading Comprehension, was reduced and the proportion devoted to the least time-consuming item type, Antonyms, was increased. The new specifications called for approximately 70% discrete items and 30% passage-based reading items. The total number of reading passages was reduced from seven to five; the Synthesis passage was dropped and only one science passage (either Biological or Physical, at the discretion of the test assembler) was included in each form. Individual passage length was specified at 400-450 words, or a total of 2,000-2,250 words for the test as a whole.

Two SAT-V subscores were also introduced in 1974: a Vocabulary subscore, based on the Antonym and Analogy items, and a Reading subscore, based on the Sentence Completion and Reading Comprehension items. Verbal test assemblers were required to produce Reading and Vocabulary subtests that had approximately the same mean item difficulty and the same standard deviation of item difficulty as the total test. Previously, Reading Comprehension items, which tended to be of low-to-middle difficulty, were balanced by difficult items of the other three types. Despite the fact that Sentence Completion items could no longer con'ain as much difficult vocabulary, because they were excluded from the Vocabulary subtest, a larger number of difficult Sentence Completion items had to be used to meet Reading subscore difficulty specifications.

In October 1978 the second of the two changes that occurred in SAT-V content specifications during the 1970-85 period was introduced: three "medium-length" (200-250 word) passages took the place of two of the "long" (400-450 word) passages. Test specifications were changed to call for a total of three long and three medium passages per final form. Each long passage had five items associated with it, and each medium passage had two to four items. Since the new specifications called for six passages rather than five, the second science passage that had been dropped in 1974 was restored in 1978; new forms of the test contained both a Biological- and a Physical-Science passage. The test assembler decided which three reading passages were to be long and which three were to be of medium length.

Other changes in the content of the SAT-V occurred as the result of attending to the concerns of females and minority groups. For example, since December 1977 new forms of the SAT-V have included a minority-relevant reading passage. Further changes associated with test sensitivity review are discussed in the section "Changes Related to the Test-Development Process."

Table 5 shows how the November and December SAT-V forms met content specifications during the period being studied. This table displays the range of numbers of items within each content category that appeared in new SAT-V forms during periods between changes in content specifications. Data for individual November and December forms are presented in Tables B-1 and B-2, respectively, in Appendix B. There were few deviations from the specifications.

SAT-Mathematical

There was only one major revision to SAT-M content specifications during the 1970-85 period (see Figure 2). Prior to the 1974-75 testing year, there were 60 mathematical items in two separately timed sections. One section contained 25 Regular Mathematics five-choice items to be administered in 30 minutes; the other section contained 17 Regular Mathematics items and 18 Data Sufficiency items to be administered in 45 minutes. Directions for each item type with examples of items are given in Appendix A. The directions for answering Data Sufficiency items were complex and required attention to details. This item type had fixed answer choices and generally required a student to assess the relevance of certain given information for determining the answer to a question. In most cases, it was not necessary to find the specific answer to an item, but rather to evaluate the contribution of two given pieces of information to the determination of a solution.

Beginning in 1974, the four-choice Quantitative Comparison item type replaced the five-choice Data Sufficiency item type. Directions for answering Quantitative Comparison items are also included in Appendix A. Test takers could answer Quantitative Comparison items more quickly than they could Data Sufficiency items. Therefore, approximately as many Quantitative Comparison items could be used in the 60-minute SAT-M as Data Sufficiency items in the 75-minute version. Tests since October 1974 have contained 40 Regular Mathematics items and 20 Quantitative Comparison items. The 25-item (30-

minute) Regular Mathematics section continued unchanged; however, the 35-item, 45-minute section containing 17 Regular Mathematics items followed by 18 Data Sufficiency items was replaced by a 35-item, 30-minute section containing 15 Regular Mathematics and 20 Quantitative Comparison items. Figure 2 and Table 6 depict the changes. The total testing time for the SAT-M was reduced by 15 minutes.

All questions used in the SAT-M during the 1970-1985 period were classified in one of four primary content categories: Arithmetic, Algebra, Geometry, or Miscellaneous. The approximate percentages of items in these respective categories were (and still are) 30%, 30%, 30%, and 10%.

In 1969 a computer-based classification system was introduced together with a computer-assisted assembly package. The content specifications were also formalized when computer-assisted test assembly was introduced. In addition to specifying the numbers of Arithmetic, Algebra, Geometry, and Miscellaneous items to be included on the test, restrictions were placed on the number of items in subclassifications within these major areas. Computer-assisted test assembly also resulted in the formalization of various dimensions of test assembly: (1) The computer-based classification system in mathematics, still in use today, added the following ability levels adapted from Bloom's (1956) taxonomy:

0 = Recall factual knowledge

1 = Perform math manipulations

2 = Solve routine problems

3 = Demonstrate comprehension of math ideas and concepts

4 = Solve nonroutine problems requiring insight or ingenuity

5 = Apply "higher" mental processes to math

(2) The setting of the stem was controlled: Problems involving money, people, etc., were classified as concrete, while those involving only algebraic symbols, numbers, and simple geometrical figures were classified as abstract.

Table 6, which contains the content specifications in place from period from 1969 to the present, shows that there was no large shift in the primary content of the test despite the shift in item types. Because of shifts in curricular emphasis, there have been one or two fewer geometry items in forms since 1974. While the formal content specifications changed very little, there was a slight change in the flavor of questions written within the area of geometry. This change is described later in the section "Changes Related to the Test-Development Process."

Table 6 includes the setting and ability specifications used since November 1969. The decrease in the number of items specified for ability level five beginning in 1974 was due primarily to the elimination of the 18 Data Sufficiency items, all of which were classified as level five because of the complexity of this item type. Forms developed for administration from October 1974 on were assembled by hand to the specifications in effect during the computer-assisted assembly period.

Table 7 contains data that shows how well November and December SAT-M forms met content specifications from 1970 to 1984. Between each change in content specifications, the ranges of items within each content category are presented in these tables. Tables B-3 and B-4 in Appendix B show the numbers of items in each content category for each November and December SAT-M form during the period studied here. Although these test forms met ability-level specifications, the number of items classified as applying "higher" mental processes (Levels 4 and 5) decreased considerably in 1974. As was mentioned earlier, this decrease was due to the use of Quantitative Comparison items in place of Data Sufficiency items, beginning in October 1974. The numbers of items classified as Arithmetic, Algebra, Geometry, and Miscellaneous, almost always met specifications.

Changes in Format

Changes in Section Order

One of the major changes that occurred with respect to both the verbal and mathematical portions of the SAT was the decision in 1974 to "scramble" sections, i.e., to change the order of the sections from one test booklet to another. The use of scrambled sections permitted proctors to seat test takers closer together in the same testing room without encouraging the copying of answers. Although scrambling made it appear that different operational versions were being used at a particular test administration, all students at a given administration were scored on exactly the same items.

Throughout the remainder of this section of the report, the following labels will be used to refer to the various separately timed parts of the SAT. "Verbal 1" refers to the longer of two Verbal sections, initially 50 items in 45 minutes, then changed in 1974 to 45 items in 30 minutes. "Verbal 2" refers to the shorter, 40-item section that was allotted 30 minutes throughout the 15-year period being studied. "Math 1" refers to the 30-minute section containing 25 regular (5-choice) math items. "Math 2" refers to the 35-item section that was allotted 45 minutes when it included the Data Sufficiency item type and then changed to 30 minutes with the switch to Quantitative Comparison items in 1974. The "one" in Verbal 1 does not refer to the first section in the test booklet, nor does the "two" in Math 2 refer to the second section. "Variable" refers to the section of the test in which pretests and equating tests are administered.

Prior to October 1974, there were five sections of the SAT arranged in one of two ways. Only one of these arrangements was used at each administration of the test. Following is a description of the two arrangements:

	Arrangement	
Section	A	B
1	Verbal 2	Verbal 2
2	Verbal 1	Verbal 1
3	Variable	Math 1
4	Math 2	Math 2
5	Math 1	Variable

From October 1974 to September 1978, the six sections of the SAT were arranged relative to each other in the following sequence: Verbal 1, Math 1, TSWE, Verbal 2, Math 2, Variable. A particular student's test booklet could begin with any one of these sections. For example, Section 1 for some students was Math 1; Section 2, TSWE; etc. The last section in this case was Verbal 1. Thus, there were six different orderings (scrambles) of sections and all six were spiralled together (collated one after the other) at each administration of the SAT. That is, at a given test center, student one received the first ordering of sections, student two the second ordering, etc. Student six received the sixth ordering, student seven received the first ordering, and the spiralling continued in this fashion throughout the test center. (In 1977-78 only five of the six possible orderings were used.)

In October 1978 scrambling of the sections at individual administrations was stopped because of the enormous complexity of the required test booklet spiraling and the suspicion that scores obtained by students who received different scrambles may not have been strictly comparable. Two fixed arrangements were then established and each administration from October 1978 through June 1980 was assigned one of the two orderings. The location of the variable component (Section 3 or 6) determined the arrangement of the five operational sections, as follows:

	Arrangement	
Section	C	D
1	Verbal 2	Verbal 1
2	Math 2	Math 1
3	Variable	TSWE
4	Verbal 1	Verbal 2
5	Math 1	Math 2
6	TSWE	Variable

Beginning with the October 1980 administration, the arrangement containing the variable component in Section 3 was dropped and two new arrangements were added. The location of the variable component continued to determine the fixed arrangement of the operational sections, as follows:

	Arrangement		
Section	D	E	F
1	Verbal 1	Verbal 2	Verbal 1
2	Math 1	Variable	Math 1
3	TSWE	Math 1	Verbal 2
4	Verbal 2	TSWE	Variable
5	Math 2	Math 2	Math 2
6	Variable	Verbal 1	TSWE

Math 2, the 35-item section that included 20 four-choice Quantitative Comparison items, was always located in Section 5 so that a standard, tailored answer sheet containing four and five response options for Section 5 could be used at all administrations. One of these three arrangements was used at each administration from October 1980 through the 1984-85 testing year.

Changes in Item Order

There were a few changes in the ordering of verbal and mathematical item types from 1970 to 1985. These variations are shown in Figures 1 and 2. The number of items decided upon for the various 30-minute verbal and mathematical sections first administered in October 1974 seemed to cause both tests to be a little more speeded than they had been before. There were also spacing problems in fitting one of the verbal sections into the allotted number of pages.

Items were rearranged beginning with the November 1975 administration to make the test less speeded and to improve spacing. The two reading passages in Verbal 1 where moved ahead of 5 Sentence Completion and 10 Analogy items. In Verbal 2 the reading passages were moved to the end of the section. Changes also occurred in the SAT-M. In the 1974-75 testing year the 20 Quantitative Comparison items were at the end of the second mathematical section--they followed 15 Regular Mathematics items. Because test-analysis data indicated that this section was speeded, 1975-76 forms had the 35 items in this section arranged as follows: 7 Regular, 20 Quantitative Comparison, 8 Regular. This arrangement ensured that most students reached the easier and faster-moving Quantitative Comparison items; the more difficult of the Regular Mathematics items, which also took more time to answer, were placed at the end of the section.

Changes Related to the Test-Development Process

Assembly and review of the verbal and mathematical sections of the SAT remained essentially unchanged during the 1960's and early 1970's. Each form of the SAT-V or the SAT-M was assembled by a single test developer and then reviewed by two test specialists, the test-development coordinator, and two editors. After test production, one final review of the typed copy was conducted by a test development staff member who had not yet seen the test. These steps in the development of SAT final forms have remained in place to the present, but have been supplemented throughout the intervening years.

Beginning in 1970, a concerted effort to pretest reading passages relevant to minority groups was initiated. A minority-relevant reading passage was included in the March 1973 form and was included in at least one new form of the SAT-V until the 1976-77 testing year. Beginning with the December 1977 form, a minority-relevant reading passage was assembled into every form of the SAT-V, a policy that has continued to the present.

In the mid-1970's, a review of each new form of the SAT-V by minority staff members at Educational Testing Service (ETS) was begun. This "minority review" became formalized on a corporate-wide basis with the establishment of the ETS guidelines for test sensitivity review (Educational Testing Service, 1980). These guidelines required, and still require today, that a trained sensitivity reviewer look at every ETS test, not only to eliminate material that might be offensive or patronizing to females or minority group members, but also to ensure that the test represents the varied contributions of these groups to American society.

One obvious change in the SAT that resulted from the institution of sensitivity review was the virtual elimination by testing year 1977-78 of the use of the generic "he" in reading passages and Sentence Completion items. In addition, the proportion of male references to female references in the Sentence Completion items and Reading Comprehension passages changed gradually throughout the period of study. (See Tables B-5 and B-6 in Appendix B.) Cruise and Kimmel (1990) noted that in Sentence Completions, the ratio of male-to-female references dropped from approximately 8:1 in 1972-77, to approximately 2:1 in 1977-82, to approximately 1:1 in 1982-87. They also noted that in Reading Comprehension passages the male-female ratio decreased from 4.3:1 from 1972-77, to 4.0:1 from 1977-82, to 3.2:1 from 1982-87. (See Cruise and Kimmel (1990) for a more detailed review of gender and minority-group references in the SAT-V.) Tables B-5 and B-6 in Appendix B give the gender references of items in November and December SAT-V forms administered from 1970 to 1984.

At the November 1976 meeting of the Mathematics Discipline Committee (now called the Mathematical Sciences Advisory Committee), one agenda topic was the level of geometry tested on the SAT-M. The Committee reviewed 10 geometry questions that had been identified by ETS staff as representing borderline content for inclusion on the SAT. Each of the questions had appeared in recent forms of the test. There was general agreement among members of the Committee that several of the questions fell outside the implied content domain for geometry that was called "informal" geometry. These questions involved either figures that some Committee members viewed as unnecessarily complicated or certain triangle relationships that might give a distinct advantage to those who had taken a formal course in geometry. The Committee recommended that the descriptive materials provided for students who take the SAT be revised to mention the various triangle relationships that might be tested and that the word "informal" be dropped from the description of geometric concepts tested. The Committee indicated that a more accurate description would refer to those geometric concepts that are typically taught at the elementary school and junior high levels. The Committee also suggested that future versions of the SAT avoid questions like some of the more complex ones presented.

Since that time, the descriptive booklet *Taking the SAT* was revised to include a review of some of the geometric relationships tested. Although there are still difficult geometry questions on the SAT, assemblers generally have avoided some of the more achievement-based exercises that depend heavily on theorems from high school geometry.

In 1977 the College Board formed a 10-member external committee of college and high school educators and administrators to provide advice regarding SAT policy and program issues and to review each newly assembled edition of the test. These SAT Committee reviews were conducted through the mail for each test form by 3 of the 10 committee members and represented the first regularly scheduled external reviews of the SAT. Prior to the appointment of the SAT Committee, a group called the Committee of Examiners in Aptitude Testing provided psychometric advice about the test, but did not routinely review new forms of the test. This latter group was disbanded in 1972, and from 1973 to 1977 there was no external review group for the SAT.

The State of New York passed legislation requiring public disclosure of SAT operational test questions administered in New York beginning with the forms administered in calendar year 1980. In October 1980 the College Board began implementing national disclosure of the SAT at four Saturday administrations and one Sunday administration during each testing year. The effects of the disclosure of test questions on the development of the SAT were many and varied. For example, item- and test-review procedures, which had always been thorough and rigorous, were given even greater attention. Test developers tried to anticipate every possible interpretation of a math problem and every nuance of a vocabulary word to be certain that alternate approaches would not yield unintended correct answers. "Fine distinctions" doubtlessly became a little less fine, and items with particularly close distractors were generally avoided in the assembly process.

Perhaps the most important effect of disclosure was that not all new forms could be kept secure so that the items could be reused in subsequent years after the forms were retired. Additionally, items that were hard to develop (such as difficult Sentence Completions) could no longer be "borrowed" from one secure final form to another to help test assemblers meet specifications more easily. The pace of final form development increased from approximately 5 new forms per year during the 1970-78 period, to 7 new forms per year in 1979 and 1980, to 9 or 10 new forms per year from 1981 to the present. Correspondingly, pretest development increased dramatically. The total number of verbal and mathematical pretests increased from approximately 40 in 1979-80 to 75 in 1980-81 to 100 in 1981-82, and has remained at about this level until the present.

Test-development staff who worked on the SAT, of course, also increased in number in response to disclosure. The number of test developers who worked primarily on the SAT-V increased from 3 to 5 during the period 1970-77, to 8 during 1978-79, 10 during 1980-81, and 14 by 1984-85. Of these staff members, the number who assembled final forms increased from 1 to 3 per year during 1970-1979 to 5 per year by 1984-1985. Similar increases in the SAT-M development staff took place over the same period of time. In addition to an increase in the number of staff involved in the development and review of both pretests and final forms of the SAT, there was a significant increase in the number of outside item writers. Prior to test disclosure, virtually all SAT-M items were written by current or former staff. As the need for additional items increased, outside item writers (primarily high school teachers) were paid to submit sets of items. Materials were prepared to familiarize writers with the

SAT content domain, and new writers were given a training exercise before their initial assignments. Outside writers have contributed significantly to the development of the SAT since 1979.

The number of outside verbal-item writers increased less than the number of outside mathematical-item writers. Even before test disclosure, Reading Comprehension sets were routinely secured from outside writers. Antonym, Analogy, and Sentence Completion items, however, continued to be written primarily by ETS staff. The reason for this fact is that outside writers tended to use general vocabulary words (and also analogical relationships) that overlapped with what had already been written and disclosed. Such overlap could be more effectively controlled if items were written and checked by ETS staff who maintained files of previously used vocabulary.

Another very significant effect of disclosure was the publication of large numbers of tests that could be purchased by students. The opportunities for test familiarization and the study of actual test items have increased year after year from 1980 onward as the College Board published individual test forms, annual compendia of disclosed test forms, and trade books such as *10 SATs* (College Entrance Examination Board, 1988), which is now in its third edition. Increased publication of test forms also led to the reformatting of score equating sections (in January 1985) so that they looked exactly like operational sections.

Early in the 1980's, the first of what has turned out to be very few flawed items was discovered as a result of test disclosure. In response to this experience, the review process was supplemented once again. Effective with the SAT-M test forms administered in October 1984 and the SAT-V test forms administered in October 1985, two 15-member external panels of subject-matter specialists were appointed to review new tests. Five of these 15 content experts reviewed each new form of the SAT-V, and three of the five then met with the particular ETS staff members who assembled the forms being reviewed. The same procedure was followed for the SAT-M. Finally, after revisions to the test had been made on the basis of these external reviews, an internal test development review meeting for each test form was held (separately for SAT-V and SAT-M) involving three ETS test development staff members in addition to the test assembler. Thus, from 1970 to 1985 the total number of internal reviewers per SAT final form increased from five to nine, and the total number of external reviewers from zero to eight.

As the result of these changes in the test-development process, new forms of the SAT have been required to satisfy a very rigorous review process. This review process helped to ensure the fairness of the SAT for various groups and provided an extra measure of quality to the development process in the 1980-85 period.

Changes in Test-Familiarization Materials

Changes in the content of the test and the test-development process also affected the information given to students to familiarize them with the SAT before the test administration. Students who register to take the SAT are furnished with an information bulletin containing test-taking tips, sample questions, explained examples, and since 1977 a full-length test. The type of information provided to students and the way that information was provided fluctuated considerably from 1970 to 1985. The number of sample items varied considerably, the length of the sample test varied from a few items to a full-length test, and the way the booklet was distributed to students was not uniform.

Table 8 provides an overview of the contents of these booklets from 1970-71 through 1984-85. The table shows that in testing year 1970-71 the number of items having an explanation of the correct answer was 16 for the SAT-V and 17 for the SAT-M, the number of questions included in the sample test was 57 for verbal and 36 for math, and the total number of pages in the booklet was 55. In contrast, the period from 1972 through 1977 had considerably fewer explained examples and sample test questions. Also, the total number of pages in the bulletin was reduced from 55 to approximately 16.

The primary reason for the shift had to do with the way the bulletin was distributed. The booklet was mailed with each student's admission ticket in 1972-73 and 1973-74. Including the booklet with the admissions ticket ensured that all students received information about the test. To reduce the expense associated with this direct mailing, however, it was necessary to shorten the bulletin. This "direct-mail" distribution procedure contrasted with the earlier and later distribution procedure in which students were told to pick up bulletins from their high school guidance offices--a procedure that may have resulted in some students taking the test without ever having received the bulletin of information. Of course, there was no guarantee that students who received the bulletin in the mail actually read it.

During the period 1978 through 1985, the number of explained items was expanded considerably and, for the first time, a full-length sample test with item data was included in the bulletin. Also, a mathematics review section was added in 1978 to give students an idea of the types of skills and content tested in the SAT-M. Examples of explained items from *Taking the SAT, 1984-85* (College Board, 1984), are in Appendix A. These examples are typical of those included in *Taking the SAT* since the 1977-78 testing year.

Impressionistic Reviews

Because changes may occur in a test in subtle ways, members of the ETS test-development staff were asked to review a number of SAT forms from the 1970-85 period. The sample included 75-minute forms from the 1970-74 period as well as 60-minute forms. Reviewers were asked to note any differences they observed between and among these tests, either of a general or of a specific nature. They were also asked to recall any subtle or philosophical changes that may have taken place in the development of the SAT during the 15-year period being studied. Six members of the ETS test-development staff who currently work on the SAT-V and other language and literature examinations were asked to review several SAT-V forms administered from 1970-71 to 1984-85. The amount of test-development experience at ETS among these staff members ranged from 10 to 25 years. All who conducted the reviews have participated extensively through the years in the development of the v^e^rbal sections of the SAT. Each reviewer examined one or two forms of the test from each of the following periods: 1970-71, 1975-76, 1978-79, and 1983-84. In all, 15 test forms were reviewed.

Each of five members of the ETS test-development staff was asked to review five SAT-M forms. All of the reviewers, whose length of test development experience varied from 2 to 10 years, had assembled final forms of the SAT-M, but none had assembled forms before 1978. Each reviewer examined two test forms developed in the period 1970-1974, one test form developed shortly after the test changed in the fall of 1974, one test form developed during the period 1978-79, and one test form developed during the period 1983-84. A total of 16 different test forms was examined. Reviewers were asked to consider attributes such as level of complexity, concrete *vs.* abstract settings, content domains of the tests, item formats, and other factors that would provide a contrast over the period.

SAT-Verbal

The most significant differences across the SAT-V test forms that were noted by the reviewers were that the testing time was shortened from 75 to 60 minutes in 1974, and that three shorter reading passages were substituted for two longer passages in 1978. These differences reflected changes in the content specifications for the test. The changes in the number of items for the four verbal item types were also noted, as were the shifts in the order of item types within the two sections of the verbal test.

None of the reviewers thought that, overall, the items from any one of the test forms were noticeably harder or easier than the items from any of the other forms. In general, the level of diction used across the Antonym, Analogy, Sentence Completion, and Reading Comprehension items seemed comparable in all of the forms that were reviewed. Two reviewers noted, however, that the introduction of Vocabulary and Reading subscores in 1974 affected the assembly of the items by item type: there were more difficult Sentence Completions and fewer very difficult Antonyms and Analogies in the more recent forms of the test.

A question was raised by one of the verbal reviewers concerning the reading passages. She wondered if perhaps the more recent passages--particularly those that were about 250 words as opposed to about 450 words--were less subtle and intricate, more straightforward and expository, than were the passages administered in the early 1970's. The short passages in particular seemed to this reviewer to require less concentration on the part of a reader in order to understand them. Two other reviewers, on the other hand, stated that the shorter passages did not necessarily seem any easier to them--just shorter.

Several differences in "surface-level" characteristics of items between the earlier and later forms of the test were commented upon. Most of these differences can probably be attributed either to the introduction of test sensitivity reviews or to the onset of test disclosure, both of which occurred in 1980.

- o Earlier forms had many more references to males (famous as well as fictional) in the sentence completions and reading passages than did later forms of the test.
- o Some "fine distinctions" seemed finer (in the discrete items in particular) in the earlier than in the later forms of the test.
- o More reading and sentence completion items (especially those classified as "humanities" and "science") seemed to assume or reward outside knowledge in the earlier than in the later forms of the test.
- o References to real people and events seemed fewer in number in the later than in the earlier forms of the test, except for references to minority group members, which increased in the later forms.
- o Unusual or esoteric words (such as metazoic, refulgent, and rebus) appeared as wrong answer choices more often in the earlier than in the later forms, but rarely were such words used as part of the stem or key of an item in any of the forms examined.

Some final comments of a more "philosophical" nature came from one of the reviewers who has worked on the SAT-V for nearly 25 years. She recalled that in the 1970's test developers tried to build the "fairest test possible for bright, informed, and motivated students." There was less concern then that a particular sentence structure or paragraph of a reading passage might be too complex or too challenging, or that the level of diction in an item might be too high. The rewarding of outside knowledge was also less of an issue. Test developers assumed that most students had studied and read widely diverse verbal material like that represented on the SAT.

"Scholastic preparedness" seemed, according to this reviewer, to take on a more diminished meaning after the lengthy SAT score decline, and particularly after a television news journal's report of a high school student who did not know the composer Wagner. Neither the student nor the journalist realized that the Sentence Completion item referring to Wagner could be answered using logic and verbal ability rather than by knowing who Wagner was or what his operas were like. The student audience for the SAT seemed different from before, according to this veteran test developer, and this perception may have subtly influenced the nature of the test. In developing the more recent forms, reading passages were less likely to be "dull" and scholarly, or technical and complex. Item writers sought "interesting" and "relevant" material. Because there were fewer references to actual historical figures (like Wagner), the test may have become less concrete, may have included less of an achievement "load" than it once did. Nevertheless, these developments in many respects represented an appropriate and desirable evolution of the test.

SAT-Mathematical

One clear contrast noted by all reviewers was the shift from Data-Sufficiency items to Quantitative Comparison items that took place in 1974. Most reviewers commented that Data Sufficiency items were more complex and required more sophisticated analysis than Quantitative Comparison items, which had more straightforward directions and required less time to answer. One reviewer commented that skills measured by Data Sufficiency items seemed more closely related to aptitude than the skills measured by Quantitative Comparisons. Another reviewer noted however, that, although .he Quantitative Comparison item types seemed easier, they frequently required more manipulation to determine the answer than Data Sufficiency items did. One reviewer indicated that the breadth of knowledge tested in Quantitative Comparison Geometry items seemed greater than that tested in Data Sufficiency Geometry items. In contrasting the 1979 and 1984 forms, one reviewer noted that the Quantitative Comparison items in the 1979 form could be solved more quickly than the 1984 Quantitative Comparisons, which contained more verbiage.

In terms of test content, reviewers commented that the forms developed before 1974 contained a few more Geometry items than later forms. One reviewer noted that the earlier forms contained Geometry items that would now be considered too achievement-like. However, another reviewer commented that Geometry items in the more recent forms had more complex figures and required greater manipulation to find the correct answer. Yet another reviewer indicated that the achievement level required to solve the Geometry items was the same across forms.

It was observed that the Arithmetic and Algebra items were somewhat more straightforward in earlier forms but required more insight in later forms.

Concrete (*vs.* abstract) settings in earlier forms tended to be in the area of Arithmetic (ratio, percent, measurement), whereas concrete settings in later forms were better balanced across content areas. Percentage, average, and age problems were more frequently tested on earlier forms of the test. One reviewer indicated that the later forms seemed to have somewhat fewer items with concrete settings.

Items appropriate for the SAT that cannot readily be classified as Arithmetic, Algebra, or Geometry are classified as "Miscellaneous." Actually, the Miscellaneous category is quite well defined in the sense that there are five subcategories with enumerated topics. One reviewer noted that the number of Miscellaneous items increased slightly between 1970 and 1976 but remained fairly constant after 1976. That reviewer observed that the more recent versions contained more "newly defined operations" than did earlier forms. Items of this type provide a definition of an abstract symbol and require the student to use and/or apply this definition.

Most reviewers made general observations about the various tests that they examined. One reviewer noted that the inclusion of more concrete items on earlier forms made these tests more interesting. Another reviewer commented that a greater breadth of general knowledge was assumed in the earlier forms (e.g., the number of days in July). Yet another reviewer observed that questions in the earlier tests were stated briefly and that there was "less concern to spell everything out." It was noted that this brevity was made possible by expecting students to "catch on" to the idea and by the use of more sophisticated language. Along this line, another reviewer noted that the number of items with more than three lines of text ranged from 5 such items in a form developed in 1973 to 14 items in a form developed in 1984.

Some reviewers offered summary comments. One noted, "With the exception of the [Data Sufficiency] items . . . , most of the items on these tests would not look out of place on today's test." Another reviewer summarized as follows: "While there do appear to be some differences between the early versions and the more recent versions of the SAT-M, . . . shifts in some characteristics of the tests seemed to be balanced or canceled by shifts in other characteristics . . . . While the introduction or additional emphasis over the years of new content areas (e.g., newly defined functions, probability, counting) may have increased the achievement level slightly, I think that this increase was balanced by not including so many arithmetic or algebra items that involved straightforward manipulation in the test."

Validity Trends and Changes in Test Content and Format

This section of the report focuses on the relationship of changes in the test itself to changes in predictive validity. Table 9 notes periods when validity changed and the SAT itself underwent little or no change. From 1970 to 1974, when validity increased, the content and format of the SAT remained constant. The only change of note was the introduction of a minority-relevant Reading Comprehension passage in one form of the SAT per year in 1972-73. The SAT booklet at that time consisted of three 30-minute sections, one of which was used to administer pretest and equating items, and two 45-minute sections. The 90 verbal and 60 mathematical items were administered in a fixed item order and in one of two fixed section orders.

From October 1974 to September 1978, when validity started to decrease, the shortened SAT was offered in a fixed sequence of 30-minute sections but in six different orders at each administration. During this time few changes occurred in the content of the SAT. The only change of any significance, in November 1975, was the relocation of reading passages and Quantitative Comparison items in an attempt to reduce the speededness of the 45-item SAT-V section and the 35-item SAT-M section. The change in the nature of the Geometry items was less important because only a few of the Geometry items used in the test had actually required any formal knowledge of geometry to answer the items.

During each of these two periods, the test remained very stable in its content, format, and statistical specifications. In contrast, the period from October 1978 to December 1981 was a time in which the content of the test changed somewhat. In October 1978 two changes occurred in the test. Shorter reading passages were introduced, thus allowing a second science passage to be used in addition to the five passages already in the SAT. In addition, the number of section orders was reduced from six to two, with only one offered at a given administration. In 1980 the testing program was required by New York

State law to begin disclosing four Saturday-administration and one Sunday-administration form of the test each year. The disclosure requirement caused the program to produce as many as ten new forms a year to replenish the pool of usable test forms. In addition, the test forms were undergoing change because of the implementation of test-sensitivity guidelines in 1980. The 1978-81 period, then, was a time when the SAT and its test-development procedures changed somewhat--a time when test forms may have been less parallel to one another than the test forms constructed in other periods.

In January 1982 the number of difficult items required on the SAT-V was reduced, but no changes were made in the content and format of the test. Thus, the 1982-85 period was also a time of stability in the SAT content and format.

During each of these four periods, however, predictive validity was in flux. Validity increased from 1970 to 1974 and decreased during each of the other periods. Except possibly for the 1978-81 period, one cannot point to changes in the test that were substantial enough to affect validity. The fact that validity increased during the early 1970's and decreased from 1975 to 1985 suggests that factors other than changes in the test were at work in causing the decline.

Still, because SAT predictive validity began to decline the very first year scores from the shortened SAT were used for validity studies, it is important to assess how changes related to the test might have affected the ability of the SAT to predict college freshman grades. The shortening of the timing of the SAT-V and the SAT-M from 75 to 60 minutes each in October 1974 was the single most important change in the SAT during the period under study. As Table 9 shows, test developers made a number of significant changes in the SAT at that time. Changes were made in the numbers of items of each item type represented in the test, the SAT-M item types, and the orders in which the sections and items were administered. Of crucial concern is the differential validity of the various item types. In the case of the SAT-V, the number of discrete items was increased at the expense of Reading Comprehension items, thus permitting 85 items to be given in 60 minutes. In the case of the SAT-M, 20 Quantitative Comparison items replaced 18 Data Sufficiency items and 2 Regular Mathematics items, thus permitting the administration of 60 items in 60 minutes instead of 75 minutes. If Reading Comprehension items were more valid than the other verbal item types, and if Data Sufficiency items were more valid than the other mathematical item types, then this change would presumably have affected predictive validity.

Internal Structure of the SAT

A number of correlational analyses and exploratory factor analyses of the relationships among the various SAT-V and SAT-M item types were conducted prior to the shortening of the test. These analyses showed that the respective verbal and mathematical item types were so highly correlated that SAT-V and SAT-M each measured essentially one primary dimension.

It is important, however, to find out to what extent the internal structure of the SAT may have changed over time. To this end, the correlations among the item types from the SAT-V and the SAT-M were reviewed across several years. The available data on item types is limited and is not available for all of November and December test forms used in the study. However, data are available for one test form from 1971 and for the November and December test forms from 1981 to 1984. Such data began to be reported routinely in test-analysis reports in 1981. Data from a December 1970 form were also available on Quantitative Comparison items, which were administered in the 30-minute variable section of that form.

The more recent data could be compared with the data from 1971 to determine any of the changes in test content and format affected what the test was measuring. Caution should be taken when generalizing to the test forms because of the limited amount of data. Apparent changes in the test items may be specific only to the forms reviewed here, or they may reflect actual variations in what the item types measured. Nevertheless, the data provide some evidence about the stability of the internal structure of the SAT.

Because the number of SAT items representing each item type changed in 1974, the unadjusted correlations between item types can present a misleading picture of changes in internal structure. To overcome this problem the correlations among item types were corrected for attenuation. This correction adjusts the correlations to reflect infinitely long tests. Any differences in the corrected correlations over time would indicate a change in what the SAT was measuring.

Table 10 gives the correlation among SAT-V and SAT-M item types for the 1971 form and the ranges of the correlations for the 1981-84 November and December forms. The corrected correlations among the SAT-V item types across these forms from 1971 were slightly high relative to the ranges observed from 1981-1984. The corrected correlations for 1971 were above or near the high ends of the ranges for the corrected correlations from 1981-84. The correlations of Reading Comprehension with Sentence Completions and with Analogies were .04 to .05 higher than the maximum values of these correlations from the 1981-84 period.

The March 1971 corrected correlations between Regular Mathematics and the various verbal item types all exceeded the ranges observed in 1981-84. The most noticeable difference in Regular Mathematics correlations for the two periods was the difference in the corrected correlations with Analogies. The 1971 value was .84, .12 higher than the maximum for the 1981-84 period. The reason for this difference is not clear--apparently the later Analogy items tapped less mathematical reasoning than the earlier items. This decrease in the overlap between the SAT-V and the SAT-M is consistent with test developers interest in reducing the overall correlation between the these two tests. Reducing the degree of overlap between the SAT-V and the SAT-M would not necessarily affect their individual validities, but presumably would provide more opportunity for the combined tests to correlate highly with a multidimensional criterion like college freshman grades.

To determine whether the switch from Data Sufficiency to Quantitative Comparison items in 1974 had an effect on what the SAT-M was measuring, the correlations of Data Sufficiency and Quantitative Comparison items with each other and with other SAT-M and SAT-V item types were analyzed. The data from a December 1970 form and from the 1971 form indicate that Quantitative Comparison items correlated about the same with Data Sufficiency items as did the Regular Mathematics items. However, the data show that Regular Mathematics items correlated more highly with Quantitative Comparison items in the forms administered from 1981-84 than with Data Sufficiency in 1971 (see the corrected correlations). It would appear that the SAT-M became more homogeneous after the switch to Quantitative Comparison items. Such a change could have an impact on validity if Data Sufficiency items captured variance in college freshman grades that was not related to Regular Mathematics. However, the relationship to SAT-V item types was quite similar for Data Sufficiency and Quantitative Comparison items. It would not appear that replacing Data Sufficiency items with Quantitative Comparison items affected much of what the SAT-M was measuring.

Although there were some shifts between 1971 and 1981-84 in correlational patterns, the differences observed were small relative to the sizes of correlations. Differences of this kind, given the form-to-form variability in correlations would not be expected to have more than a negligible effect on the predictive validity of the test. On the other hand, one would like direct evidence of the predictive validity of the different item types.

Special Studies

Unfortunately, data available through the Validity Study Service did not permit analyses of the validity of the various item types; the records contain only reported SAT scores. There is, however, some direct evidence of predictive validity of the different item types from special studies.

Schrader (1973). Schrader studied the validity of the Quantitative Comparison item type before it was used operationally in 1974. A 30-minute, 55-item Quantitative Comparison test was administered to students at 12 colleges, including the three service academies, in the fall of 1970. Sample sizes ranged from 91 to 987 students. Scores on this test were correlated with freshman grades collected at the end of the second semester. These correlations were compared with correlations based on scores from the 75-minute SAT-M, which were available from student records at the colleges. In addition, the multiple correlations of the high school record, the SAT-V, and the Quantitative Comparison test were compared with the multiple correlations of the high school record, the SAT-V, and the SAT-M. Schrader found that despite its shorter length the Quantitative Comparison test, singly or in combination with the high school record and the SAT-V, had higher validity coefficients than the SAT-M for about half of the groups studied. He also found a tendency for the Quantitative Comparison test to perform better than the SAT-M, then composed of Regular Mathematics and Data Sufficiency items, for groups having relatively high mean SAT-M scores. Schrader concluded that there was no marked tendency for the SAT-M to be any more valid than the Quantitative Comparison test. Based on the results of this study, the testing program expected no losses in predictive validity from the substitution of Quantitative Comparison items for Data Sufficiency items.

Schrader (1984). In another study Schrader provided additional evidence of the predictive validity of item types. In a study of SAT-V item types, Schrader used matched student records from the December 1980 SAT files and the Validity Study Service files for the entering class of 1981. His final sample consisted of 11,320 students from 48 colleges, each of which had 95 or more students with complete data. He analyzed data for eight-item subsets matched in difficulty as well as for subsets based on all available items. His basic analyses provided validity coefficients and multiple correlation coefficients for each of the 48 colleges. Although Antonyms and Analogies had the highest validities in about twice as many colleges as the other verbal item types, the median validities differed only slightly. Schrader concluded that the verbal item types have similar validities and that changing the mix of items in the SAT-V was unlikely to affect predictive validity.

Burton, Morgan, Lewis, and Robertson (1989). Burton and her cohorts, in one part of their study, investigated the predictive validity of item types in the SAT and the Test of Standard Written English (TSWE). They used pooled data on about 49,000 students from 196 colleges. These data resulted from matching the November 1984 SAT and TSWE scores with college freshman grades from the Validity Study Service for the entering class of 1985. The researchers computed validity coefficients for the full sample after first adjusting the freshman grades at each college to reflect differential selection on high school grades. Using reliabilities available from the November 1984 test-analysis sample, the researchers estimated the validity coefficients for full-length (60-minute) tests of each item type. For the SAT-V, the most valid item type for the total group was Reading Comprehension (r = .44). The estimated validity of a test made up of all Reading Comprehension items was, however, the same as the existing verbal test, which was composed of a mixture of item types. For the SAT-M the Quantitative Comparison item type (r = .44) was more valid than Regular Mathematics items (r = .41) for the total group. Yet the existing SAT-M, with its mixture of items, was estimated to have a validity only .01 lower than that of the Quantitative Comparison test.

Conclusion

These analyses all point to the conclusion that changing the mix of item types in the SAT in 1974 had little effect on the predictive validity of the SAT. Although a slight decrease was evident in the overlap between the SAT-V and -M when the shortened SAT was introduced, there is no evidence that what the tests measured was substantially altered by changes in the mix of items. Since the validity evidence suggest strongly that the various SAT-V or -M item types correlate similarly with college grades, the shortening of the SAT in 1974 and the slight content changes thereafter probably had little effect on the ability of the SAT to predict college grades. The question of test length and its possible effect on validity is addressed in the reliability section in the next part of the report.

4. CHANGES IN STATISTICAL CHARACTERISTICS

After the test is administered and scores are reported, the statistical characteristics of the test are analyzed. The results of this analysis are contained in a test-analysis report, which is prepared for each new form of the SAT. The descriptive statistics in this report are useful for evaluating the extent to which the SAT met statistical specifications, the difficulty and speededness of the test for the group, the internal-consistency reliabilities of the various test scores, and correlational patterns among the various SAT and Test of Standard Written English scores. A good example of a test-analysis report is the one prepared for the December 1984 SAT form (D. Wright, N. Wright, & Weber, 1985). Two reader's guides (Walker, 1981; Educational Testing Service, 1989) explain many of the terms and concepts referred to in SAT test-analysis reports.

All item and test statistics, particularly test difficulty and speededness, are influenced to some degree by the ability of the test takers. The less able the group, the more the test will be difficult and speeded for that group. When the test is not appropriate for a group of a certain ability, the reliability and correlations with other variables will be lower than if the test were appropriate. Whenever the ability of the average SAT test taker changes, the statistical characteristics of the test will change. Therefore, before proceeding with the assessment of the statistical characteristics of the SAT, trends in the average ability of the test takers are reviewed.

Pronounced trends in SAT scaled scores are apparent in November and December for graduating classes from 1971 to 1985. The mean scaled scores on the November SAT-V and SAT-M for graduating seniors tended to decline from 1971 to 1982, and then increased (see Tables 2 and 3). For December administrations, the drop in senior means was more substantial. From 1971 to 1980, average SAT-V and -M scaled scores for December test takers decreased by approximately one-half of a standard deviation; in 1985 the average scaled scores increased slightly. While a decrease in average scaled scores of the size observed for the November administrations would have had a slight effect on some item and test statistics (e.g., speededness and reliability), a decline of the magnitude found for the December administrations could have had a noticeable effect.

Test-Analysis Samples

Ideally, perhaps, for the purpose of relating item and test data to SAT predictive validities, one would analyze item and test data for the validity-study samples. Such data, however, were not available because the item response of individual students are not routinely saved on the VSS data base. The item and test data used for this study came instead from test-analysis samples who were tested at November and December administrations from 1970 to 1984. Each of these samples consisted of approximately 2,000 test takers. Prior to 1981 these test-analysis samples were drawn to be statistically representative of the total populations of November and December test takers. Beginning in 1981, because of the increasing numbers of junior high, sophomore, and adult test takers, these samples were restricted to include only high school juniors and seniors.

The test-analysis samples did not represent either the validity-study samples or the November and December senior cohorts. Nevertheless, as Table 11 shows, the means of the November and December test-analysis samples were very similar to the means of the November and December seniors. The test-analysis samples, like the November and December seniors, scored lower on the SAT on average then did the validity-study samples. This difference is to be expected, as those enrolled in college are a more select group than those who applied for admission. Although the test-analysis samples were somewhat less able than the validity-study samples, they should be adequate for purposes of this study, which focuses primarily on test-related characteristics and their interaction with test takers, not on test takers themselves.

Table 11 gives the sample sizes, scaled-score means, and scaled-score standard deviations of the November and December test-analysis samples that provided data for this study. The item- and test-analysis data were taken from test-analysis reports for the November and December forms administered in the years 1970 to 1984.

Changes Related to Statistical Specifications

Statistical specifications govern the distribution of item difficulties and the average correlation of items with the total test in each SAT test form. The SAT statistical specifications are not geared to the average test taker. Rather, they are intended to provide relatively more measurement power in the middle-to-upper part of the score range, the range of interest to those making admission decisions at user institutions. Thus, the SAT is by design somewhat difficult for the average test taker.

Item Statistics

The specifications are expressed in terms of two item statistics. (1) Item difficulty. The difficulty of an item can be expressed as the proportion or percentage of test takers who answer the item correctly. For the SAT the proportion correct is computed by dividing the number of test takers who obtain the correct answer by the number of test takers who reach the item. This procedure in effect treats an omission as a wrong answer. ETS transforms the proportion correct into "delta," which is used as the primary measure of item difficulty. The ETS delta scale is a nonlinear conversion of the proportion correct. The proportions correct are transformed to a normal deviate with a mean of 13 and a standard deviation of 4. Deltas are inversely related to the proportion correct. For example, an item with a delta value of 17 is equivalent to a correct answer rate of 16% in the analysis group, a delta of 13 corresponds to 50% correct, and a delta of 9 represents 84% correct.

Raw (observed) item deltas, the deltas calculated on the analysis group, furnish an estimate of the relative difficulty of the item for the group on which the analysis is based. These deltas are influenced by the ability level of the analysis group. When a group is very able, the average percent correct for all items on the test is higher than when the identical items are administered to a less able group. Therefore, observed deltas are not an ideal measure of the difficulty of a test. To remove the effects due to differences among groups in ability levels, ETS equates observed delta values to a delta scale based on the performance of a common reference population. Item-difficulty specifications for the SAT are expressed in terms of these equated deltas.

(2) Biserial correlation. The correlation of the item response (right *vs.* wrong) with the total test score is used at ETS to measure the degree to which the item discriminates high-ability from low-ability test takers. In computing this correlation, omitted items are counted as wrong. The biserial correlation, an estimate of the correlation of two normally distributed variables, is used instead of the point-biserial correlation because it is less influenced by the difficulty of the item. Still, the biserial correlation is affected by the ability of the group of test takers on which it is computed and is attenuated somewhat when the group finds the item very easy or very hard. The biserial correlation may be interpreted like a usual Pearson correlation coefficient, which varies from -1.00 to

1.00, although theoretically the biserial correlation is unbounded in both directions.

Changes in Statistical Specifications

Equated deltas. From 1966 to 1974, the SAT-V statistical specifications called for a mean equated delta of 11.7 with a standard deviation of 2.9. (See Table 12, which gives the statistical specifications for both the SAT-V and the SAT-M.) The distribution of item deltas was unimodal and centered around a delta of 12. Specifications for the SAT-M required a mean delta of 12.5 and a standard deviation of 3.1. The delta distribution for the SAT-M was also unimodal, centering on deltas of 10 and 11. However, the distribution was skewed toward the difficult end of the scale. Tests built to these specifications measured best in the middle part of the score range.

With the shortening of the SAT in October 1974, the specifications were changed to make the test more appropriate for the then-current group of test takers, who were somewhat less able than previous test takers. Some exploratory work using item response theory helped set these specifications. The overall test was made slightly less difficult. For the SAT-V, the specified average delta was reduced from 11.7 to 11.4 and the standard deviation of item deltas was increased from 3.1 to 3.3. The delta distribution for the SAT-V was made bimodal in an attempt to lower the difficulty of the test, yet still maintain measurement power at the high end of the scale.

The statistical specifications for the SAT-M also changed. Although the distribution remained unimodal, the average delta was reduced from 12.5 to 12.2, and the standard deviation was increased from 3.1 to 3.2. The number of easy items was increased slightly, and the number of moderate items was decreased. No changes were made to the number of difficult SAT-M items included in each test. These specifications were also set to lower test difficulty and still maintain measurement power at the upper end of the scale. These revised specifications are still used for the SAT-M.

In January 1982 the specified delta distribution for the SAT-V was changed due to difficulties in writing items with equated deltas greater than 15. The new specifications were similar to those that existed before 1966 (see Donlon, 1984). Item-response-theory methods were again used to adjust the specifications. While the average item difficulty did not change, the specifications called for fewer difficult and fewer easy items for tests administered from January 1982 on. The intent was to maintain measurement power in the middle-to-upper parts of the score range. Some measurement power was lost, however, at the high end of the scale as a result of the change in specifications.

Average biserial correlation. In addition to the item-difficulty distribution, SAT statistical specifications regulate the average biserial correlation of the items with the total test. The average biserial correlation is specified in terms of pretest item statistics, but evaluated in terms of statistics available on items used in final forms. Pretest items do not contribute to the total-test score because they are placed in sections external to the operational test; but when items are evaluated in final forms, these items are included in the total-test score. To adjust the average item-test biserial correlation for this effect, it is necessary to add .05 to the specified average correlation for the SAT-V and .06 for the SAT-M.

Although a distribution of item-total correlations is not specified, test developers control the distribution to some extent by using mostly items with biserial correlations greater than or equal to .30. Biserial correlations of .20 to .29 are occasionally allowed for difficult or easy items. Items with high biserial correlations are used to ensure high reliability, but the number of such items is constrained by the average biserial correlation. This constraint helps test developers preserve the heterogeneity of the test necessary for adequate validity.

The specified average biserial correlation for the SAT-M has remained constant at .47 from August 1966 to the present. Only one modification to the specified average biserial correlation was made for the SAT-V during this time period. In 1974 the average was increased from .42 to .43. This increase was instituted because the pool of available items tended to have biserial correlations greater than .42. From January 1982 on, the specifications permitted a deviation of .02 in the mean biserial correlation, so that test developers could use more of the existing item pool.

Changes in Actual Statistics

Tables 13 and 14 give actual mean equated deltas, standard deviations of deltas, and mean biserial correlations for November and December SAT forms administered from 1970 to 1984. Ideally, the actual statistics would match the specified statistics. The November and December SAT-V and SAT-M forms met specifications for the most part from 1970 to 1984.

The actual mean deltas reflect the decrease in test difficulty in the fall of 1974 for the SAT-V and the SAT-M. They also indicate that the December SAT-V forms were easier than previous forms. The data also show that, throughout the period reviewed in this report, there were some slight deviations from specifications in the actual mean equated delta. Prior to 1974 the November and December SAT-V forms were more likely to be more difficult than specified on average. In contrast, the tests were more likely to be easier than specified for the period 1974 to 1984. However, in only one case, the November 1980 form, was the difference between the actual and the specified average delta larger than .2. Prior to 1974 the November and December SAT-M test forms were more likely to be easier than the specifications called for. From 1974 on, the test forms were sometimes harder and sometimes easier than specified. The largest discrepancies occurred for the November and December 1975 forms, which had mean deltas that were .4 lower than the specified value of 12.2, and for the November 1984 form, which had a mean delta that was .4 higher. To put these differences in perspective, an average actual delta of 11.2, which is .2 below the 1974 and 1982 specification for the SAT-V, represents an mean proportion correct of .67 in the reference group as opposed to .66 for the specified value. Therefore, the deviations from the specifications were slight.

The deviations between the specified delta distributions and the actual distributions (see Tables 15 and 16 and Tables B-7 through B-10) were relatively minor from 1970 to 1984. In genera!, the mean frequencies came within 2.0 of the specified values. Most of the time too many items at one difficulty level were balanced by too few items at a neighboring level.

The November 1978, November 1979, December 1976, December 1977 and December 1983 SAT-V forms had standard deviations of equated deltas that were .2 lower than the specified value of 3.3. Also, the November 1973 and December 1973 SAT-V forms had standard deviations that were too high by .2. Otherwise the SAT-V standard deviations were within .1 of the specified value. Several SAT-M forms also had standard deviations that deviated more than .2 from the specified value. These were the November 1973, 1974, 1981, and 1983 forms and the December 1977 and 1984 forms--which had standard deviations that were higher than specified--and the December 1974, 1976, and 1981, forms--which had standard deviations that were lower than specified. Despite these outliers, no systematic trends in delta standard deviations were evident for either the SAT-V or the SAT-M.

The mean item-total biserial correlation for the November and December SAT-V forms generally came close to the specified value. The November mean biserial correlations were higher than those for December and varied around the specified value until 1980, when they were higher than specified. The December SAT-V mean correlations, on the other hand, tended to be lower than specified throughout the 15-year period. The mean biserial correlations for the SAT-M fluctuated more than those for the SAT-V and tended to be higher than required. As for the SAT-V, the November mean correlations tended to be higher than those for December. In addition, for both November and December the SAT-M mean correlations were for the most part higher than specified. The mean biserial correlation with the largest deviations from specifications came from the November 1975 and December 1984 SAT-M form, whose mean correlations were, respectively, .05 and .04 higher than specified. The forms with mean biserial correlations higher than specified would tend to have higher reliabilities than the other forms but might lack the breadth of coverage desired in the test.

Changes in Other Measures of Test Difficulty

Changes in Score Conversions

Statistical specifications for test difficulty are stated in terms of equated deltas. Changes in test difficulty affect not only equated deltas but also score conversions, which provide another indication of changes in test difficulty. Score conversions are derived through the process of equating, which is discussed in the next part of the report. Here the discussion focuses on the results of equating, the relationship between raw scores and scaled scores as it relates to test difficulty.

If a test form is built to be easier than a previous form, then, in general, a given raw score will convert to a lower scaled score on this form than on the previous form. That is, obtaining the correct answer on an easier test does not count as much as on a harder test. Conversely, if the form is made harder, then a given raw score will convert to a higher scaled score. Changes in score conversions should ideally be consistent with changes in equated deltas. A decrease in mean equated delta should, in general, translate into a decrease in the scaled score that corresponds to a given raw score. But because the delta equating process is independent of the score equating process, it is important to check the extent to which they produce consistent results.

The primary result of equating is a conversion table that gives scaled scores that correspond to particular raw scores. Table 17 and Figure 3 give the scaled scores corresponding to the midpoints of the raw-score ranges for the November and December SAT forms administered from 1970 to 1984. Tables B-11 and B-12 in Appendix B provide conversion information for these forms at several raw scores. Tables 18 and 19 give the scaled-score ranges corresponding to selected raw scores for all new forms of the SAT administered from March 1970 to January 1985. In interpreting the data in these tables, one should remember that a less difficult, shortened SAT was introduced in the fall of 1974. In addition, the verbal statistical specifications were changed with the January 1982 form to reduce the number of difficult verbal items required on the test while maintaining the average difficulty level of the test.

Panels c and d in Figure 3 show the scaled scores corresponding to the midpoints of the raw score ranges for November and December SAT forms from 1970 to 1984. These plots may be compared with those for mean equated deltas (panels a and b) to see whether the results of score equating are consistent with the results of delta equating.

For the SAT-V both indices exhibited reasonably consistent trends. The drop in test difficulty in 1974 is apparent in Figure 3. In addition, many but not all of the large changes in one of the indices were matched by large changes in the other index. For example, the unplanned drops in the mean equated deltas for the SAT-V forms administered in November 1972 and November 1980 were matched by corresponding drops in scaled scores. However, the increase in the scaled score corresponding to the raw-score midpoint in December 1977 and the increase in mean equated delta in November 1982 were not matched by comparable increases in the other index. Both indices showed that from 1980 to 1984 many of the SAT-V forms were easier than previous forms.

For the SAT-M the planned decrease in test difficulty in 1974 was apparent in the plots for both indices. In contrast to the SAT-V results, however, there were many inconsistencies between the two indices. For example, the unplanned decreases in mean equated deltas for the November and December 1975 forms were not consistent with the scaled-score conversions. Whereas according to the score data, most of the SAT-M forms administered from 1981 on were relatively easy, the mean equated delta data suggested no such trend. This inconsistency suggests that the equated delta scale may have drifted upward relative to the score scale, which for score data is the more accurate of the two indices.

The tables of score ranges (Tables 18 and 19), which cover all of the forms administered during the period studied, provide data across the entire score range and not just for scores at the midpoint of the raw-score range. These tables show some reasonably clear patterns-- patterns that are consistent for the most part with the observations noted for the November and December forms: (1) In the period following the introduction of the shortened SAT, the scaled scores corresponding to particular raw scores decreased, especially for the SAT-V, indicating an easier test. (2) In the upper part of the score scale, the SAT-V scaled-score ranges gradually shifted downward from 1974 to 1985. (3) The SAT-M scaled-score ranges shifted downward in the January 1982 to January 1985 period--despite the fact that the statistical specifications remained unchanged.

Another observation that can be made about the score-range tables is that the variation from form to form was relatively small at any given raw score. The largest difference between the maximum and minimum value in a score range was 40 points for the SAT-V and 50 points for the SAT-M. Counts of the numbers of raw scores in Tables 18 and 19 with scaled-score ranges of a particular magnitude (see Table B-13 in Appendix B) show that the most common range was of size 30. They also show that the 1974-78 period manifested the least form-to-form variation for SAT-V, and the 1974-78 and 1978-81 periods for SAT-M, indicating that test forms were more parallel and that, thus, there was less of a burden placed on equating in these periods. These counts indicate that the shortened test forms, those introduced in the fall of 1974, were no less parallel than those given in the 1970 to 1974 period. They also indicate, however, that the SAT-M forms in the 1982-85 period were less parallel.

Changes in Relative Test Difficulty

Indices of relative test difficulty. In addition to an equated delta, an observed delta is calculated for each item on a test. The mean observed delta indicates the average item difficulty for the group that took the test and as such is a measure of the relative difficulty of the test for the group.

Another index of the test's relative difficulty is the mean raw score on the test divided by the number of test items. In the case of a test that is scored by counting the number right, this index is equal simply to the mean of the individual proportions correct of the items. The SAT, however, is scored by formula in that a fraction of the number wrong is subtracted from the number right. In this case the index is equal to the mean over items of the item proportion correct minus of a fraction of the proportion wrong. This value is referred to here as the "mean adjusted proportion correct" in the remainder of this discussion. Because omitted questions are not counted wrong in calculating the mean adjusted proportion correct, this index is not distorted by differences in omitting patterns across groups. There, the mean adjusted proportion correct is for most purposes a better measure of the relative difficulty of a formula-scored test than the observed delta is.

To provide optimal measurement power for a group, a test should be of middle difficulty, the value that corresponds to a score halfway between a chance score on the test and the maximum possible score. The middle-difficulty equated delta for SAT-V test forms and SAT-M test forms administered prior to October 1974, tests that consisted of five-choice items, is 12.0, which corresponds to 60% correct. The corresponding value for the SAT-M given from October 1974 on, a test composed of 40 five-choice items and 20 four-choice items, is 11.9, which corresponds to 61% correct. (See Educational Testing Service, 1989, for more detail.) In terms of the mean adjusted proportion correct, .50 is at middle difficulty for both the SAT-V and the SAT-M.

Trends. Given that the mean equated deltas specified for the SAT-V and the SAT-M were lowered in 1974, the relative difficulty of the test would be expected to drop somewhat in the absence of changes in the ability of the test takers. The ability of test takers, however, did not remain stable over the years. The average test scores declined in the mid-to-late 1970's and increased from 1980 to 1984 on both the SAT-V and the SAT-M. The decrease was larger for the December administrations than for the November administrations. Since observed mean deltas and mean adjusted proportions correct reflect trends in average test-taker ability as well as the changes in the difficulty of the test, the tests should have become relatively more difficult in the mid-to-late 1970's and easier in the early 1980's.

The trends in these two indices were reasonably consistent for November and December test forms from 1970 to 1984 (see Figure 4 and Table 20). Because the mean adjusted proportion correct does not treat omitted responses as wrong answers, it is the more accurate of the two indices. Therefore, reference is made primarily to the mean adjusted proportion correct in the following discussion. The mean adjusted proportions correct obtained by the SAT-V test takers ranged from .40 to .44 for November test takers and from .36 to .39 for December test takers from 1970 to 1984. Therefore, the test was relatively difficult for both populations, but particularly for the December test takers. The mean adjusted proportions correct for November SAT-V forms increased from about .41 in 1970-73, to .43 in 1974-79, to .44 in 1980-84, indicating that the test became easier for the groups. No such increase was observed for the December forms, probably because of the steeper decline in total-test performance for December test takers. The increase in the mean adjusted proportion correct in 1974 reflected the planned decrease in test difficulty of the SAT-V forms. Except for the years 1977 through 1979, when the mean adjusted proportion correct fell to a low of .36, this statistic was relatively stable for the December test takers throughout the 1974-84 period. Presumably, the decline in the average ability of the test takers accounted primarily for the decrease in mean adjusted proportions correct from 1977 to 1979 (see Table 11). The three-year increase in the December mean adjusted proportions correct from 1982 to 1984 was due to an unintended reduction in test difficulty coupled with an increase in the average ability of the test takers.

For the SAT-M the decrease in test difficulty effected in 1974 was not evident in relative test difficulty in either November or December. Although the SAT-M data did not show a decrease in relative test difficulty in October 1974, if the specified mean delta had not changed, the test would have been even more difficult for the test takers. The mean adjusted proportions correct for November test takers showed a slight decrease from 1974 to 1980 relative to the 1970-73 period and then a noticeable increase from 1980 to 1984. The increase was probably due to both a decrease in real test difficulty and an increase in the ability of the test takers. For December administrations of the SAT-M, the mean adjusted proportions correct showed first a decrease and then, after a large increase and decrease in 1980 and 1981, a substantial increase from 1981 to 1984. The decrease in mean adjusted proportion correct from 1971 to 1977 corresponded to the decline in average ability for the December test-taking population during this period. The upswing from 1981 to 1984 was due to an interaction of test difficulty and test-taker ability.

To summarize, the observed mean delta and the mean adjusted proportion correct demonstrated similar patterns for November and December SAT-V and SAT-M test forms. The trends in relative test difficulty occasionally followed patterns expected from changes in average test-taker ability but in general reflected changes in test difficulty as well as changes in the test-taking population. The November forms of the SAT-V and the SAT-M tended gradually to become less difficult for test takers from 1970 to 1984. The December forms became slightly more difficult in the mid-1970's and then decreased in difficulty to the point that the forms were about as easy for the test takers as tho. e offered in the early 1970's. Thus, the more recent forms measured the ability of the average test taker as well as or better than the earlier forms. Still, the SAT remained difficult for the average test taker and presumably for the average student in validity-study samples. The closer a form is to middle difficulty for a group the better its measurement power for that group.

Changes in Speededness

Speededness may be defined as the extent to which test takers are unable to complete a test section within the time allotted. Data for assessing the degree of speededness are based on the unanswered questions after the last answer marked by each test taker and do not take into account previously omitted items. Such unanswered items are said to be "not reached."

Speededness Indices

Indices that are used in determining the degree of speededness of a test section of the SAT include the percentage of test takers completing 75% and 100% of the section. The percentage of test takers completing 100% of the test section, however, is affected by the presence of a very difficult question at the end of the test. Thus, it is impossible to tell whether a test taker meant to omit the item or did not have time to consider it. Another approach to judging speededness involves examination of the variance of not-reached items compared with the SAT-V or the SAT-M test section formula-score variance. The mean and standard deviation of the number of items not reached by the group provides additional information useful in interpreting the other speededness indices. All of these indices except the mean and standard deviation of the number not reached take into account the length of the test. The latter indices can be adjusted for test length by dividing by the number of items.

It has long been the practice of ETS to regard a test as essentially unspeeded if virtually all of the test takers complete 75 percent of the test items (Swineford, 1974). This criterion is arbitrary and not rigidly applied. Also, it has been suggested (Swineford, 1974) that a variance ratio exceeding .25 indicates that a speed factor is probably present. Generally, the larger the variance ratio is, the larger the mean number of unreached questions relative to the total number of items in the section.

When judgments about the relative speed of various test forms are to be made, the ability levels of the groups, as defined by their scaled-score means, need be taken into account. Otherwise, test forms taken by less-able groups may mistakenly be evaluated as more speeded than test forms taken by more-able groups. Test-analysis reports give the data for the sample at hand: No adjustment for ability level is made in the speededness indices. The mean scaled scores on the test-analysis samples (see Table 11) permit an informal assessment of the effect of sample. Because of the lower average ability levels of the December samples, the test sections would be expected to be more speeded for these samples.

Expected Changes

Changes in test format, item types, and time limits like those that occurred to the SAT in the fall of 1974, might be expected to decrease speed factor in some test sections and increase it in others. In the shortened test, 85 SAT-V items and 60 SAT-M items were administered in 60 minutes, whereas in previous forms 90 SAT-V items and 60 SAT-M items had been administered in 75 minutes. Although fewer reading passages were used in the SAT-V and Quantitative Comparison items replaced Data Sufficiency items in the SAT-M, test developers recognized that these changes could still cause the new 45-item verbal section and the 35-item mathematical section to be more speeded. In fact, in response to a concern about speededness, the formats of these two sections were revised in 1975 (see Figures 1 and 2).

In the fall of 1978 another change occurred that could have affected speededness: the number of reading passages was increased from five to six. This change might have affected the speededness of the 40-item Verbal 2 section, in which two long passages of 400-450 words were replaced by three medium passages of 200-250 words (see Figures 1 and 2).

Speededness is affected by the ability of the test takers as well as by changes in the test. The November and December scaled-score means of the test-analysis samples decreased steadily from 1970, reaching a low during the 1979-81 period, and then rose somewhat from 1982 to 1984. Moreover, December means for both the SAT-V and the SAT-M decreased more sharply than November means, about 25 and 30 scaled-score points versus about 45 and 55. Any decrease in ability was expected to cause the test to be more speeded.

Trends

Tables 21 to 24 and Figures 6 and 7 provide data on speededness indices for November and December test forms administered from 1970 to 1984. The data on the three speededness indices considered here--percentage of the group reaching 75% of items, variance ratio, and mean number of not-reached items--tended to exhibit reasonably consistent patterns throughout the 1970-84 period. That is, with minor exceptions, as the percentage reaching 75% of the items decreased, the variance ratio and mean not reached increased.

Only 5 of the 120 sections administered in November and December from 1970 to 1984 would be considered unspeeded if the rule that virtually all (99.8% or more) of the test takers had to reach 75% of items were applied strictly. However, since most of the percentages exceeded 98%, application of this rule is misleading. In general, the sections appeared to have only a slight amount of speededness.

As expected, all four SAT-V and -M sections were more speeded--although only slightly--for December test takers than for November test takers across the 15-year period. The speededness indices frequently changed in ways that were inconsistent with changes in test-analysis sample means. Apparently, changes in test format, difficulty, and timing in 1974 and in item-type location in 1975 affected speededness in ways that were not accounted for by changes in group ability.

The separately timed sections were expected to be differentially speeded. In terms of the percentage of test takers completing 75% of the test, the longer Verbal 1 section appeared more speeded than Verbal 2 for November test takers from 1975 on. This section also appeared more speeded than Verbal 2 in December except for the forms administered in December 1974 and December 1984. The Verbal 1 sections in the newly introduced 1974 forms--November in particular--were unexpectedly among the least speeded of the Verbal 1 sections. The Verbal 1 section appeared to be more speeded after the major specifications changes in 1974, although not in 1974 per se. Although subject to considerable fluctuation, this section tended to become increasingly more speeded from 1974 to 1984 for November test takers and from 1974 to 1983 for December test takers, despite the increase in test-taker ability from 1981 to 1984 for November and 1982 to 1984 for December. After 1975 there was more variability in the indices for Verbal 1, suggesting speed factors in specific forms--perhaps due to this section being more difficult in some forms than others.

The 40-item Verbal 2 section tended to be unspeeded for most test takers throughout the 15-year period. The only exception was the November 1974 form, which was considerably more speeded than other Verbal 2 sections--perhaps because it was a relatively difficult section and the Reading Comprehension items were in the middle of the section. (In 1975 the 15 Reading Comprehension items were moved to the end of the section in an attempt to reduce speededness.) The addition of shorter Reading Comprehension passages and other changes in the SAT-V in 1978 apparently did not affect the speededness of either verbal section.

Relative to the Math 1 sections given in the 1970-73 forms, the 25-item Math 1 section became relatively unspeeded with the introduction of the shortened SAT in 1974, presumably because of the decrease in test difficulty. Exceptions were the Math 1 sections administered in November 1976 and December 1981. The 35-item Math 2 section was relatively unspeeded throughout most of the 15-year period except in November 1974 and December 1972-75. In the 1974 forms the sections were particularly speeded. Since this section contained a new item type, Quantitative Comparison, test takers were unfamiliar with the format and may have worked more slowly. Apparently, moving the Quantitative Comparison items from the end to the middle of the section in 1975 had the desired effect of making the section less speeded. This change allowed test takers to reach these items before they answered the difficult Regular Mathematics items at the end of the section.

To summarize, the shorter section of the SAT-V and the longer section of the SAT-M tended to be relatively unspeeded for most test takers for November and December forms administered from 1970 to 1984. These sections became temporarily more speeded when the shortened SAT was introduced in 1974, but the change in the ordering of the items within sections in 1975 reduced speededness to previous levels. The longer SAT-V section gradually became more speeded for test takers. The shorter SAT-M section, on the other hand, gradually became less speeded, and was relatively unspeeded for test takers from 1976 on. The addition of a reading passage and the shortening of the reading passages in the shorter SAT-V section in 1978 seemed not to make the SAT-V more speeded.

Changes Related to Reliability

The reliability of SAT scores is assessed in a number of ways. The reliability data available for this study included internal-consistency reliability estimates and test-retest correlations from spring of the junior year to fall of the senior year.

The internal-consistency estimate of reliability for a SAT-V or SAT-M test form assesses the extent to which the items in the test form measure the same underlying factor. This estimate does not take account of differences among test forms and thus does not include the effects of equating.

The test-retest estimate of reliability assesses the degree to which a second test administration yields similar scores for the same individuals. The test-retest reliability estimates, which in the case of the SAT are based on scores on alternative forms of the test, provide a measure of the stability and equivalence of the test scores across different administrations and thus take account of differences among test forms. Unlike the internal-consistency reliability estimate, this estimate is not inflated by test speededness. Because the test-retest scores came from spring and fall administrations, however, the reliability estimate is somewhat attenuated because of real changes in the test takers that occurred over time. Ideally, test-retest scores would come from administrations only a few days or a few weeks apart.

Internal-Consistency Reliability Information

Indices. Several indices can be used to assess the internal consistency of a test. The most widely used of these is coefficient alpha, which reduces to Kuder-Richardson Formula 20 (KR20) when test items are scored dichotomously. However, for the SAT, a formula-scored test, the test taker's response to an item is scored trichotomously: 1 for the right answer, 0 for no answer, and -1/(number of response alternatives minus 1) for a wrong answer. The Dressel (1940) adaptation of KR20 for formula-scored tests is used to calculate an internal- consistency reliability coefficient for the SAT. (See Educational Testing Service, 1989, for details.) The Dressel adaptation, which is also equivalent to coefficient alpha, is applied to each separately timed section of the SAT-V and of the SAT-M.

The standard error of measurement for the section is calculated from the reliability coefficient in the usual way:

$$SEM = s_x \sqrt{(1 - r_{xx})},$$

where s_x is the standard deviation of section scores and r_{xx} is the reliability of the section. The standard errors of measurement of the sections are used in turn to calculate the reliability for the entire test as follows:

$$r_{tt} = 1 - \frac{\sum SEM^2}{s_t^2},$$

where s_t^2 is the variance of the scores on the total test, and SEM^2 is the square of the scaled-score standard error of measurement. The SEM is a function of the reliability of the test and the standard deviation of test scores. The SEM, however, depends less on the particular group that takes the test than does the reliability coefficient and thus is more comparable across administrations.

Reliability coefficients are influenced by the variability of the test takers who happen to be taking the particular test form. Other things being equal, the larger the score variance is, the higher the test reliability. To provide a purer measure of test reliability, the reliability coefficients were adjusted to reflect a standard deviation of 100, so as to reduce the effect of group variability over time. The formula used to make this adjustment is:

$$1 - \frac{SEM^2}{100^2},$$

where 100 is the assumed standard deviation of a common reference population. This formula assumes that the standard error of measurement is constant across different score levels and is thus unaffected by changes in total-test-score variation. The adjusted reliabilities differ from the standard error of measurement only in scale. Since the standard deviations of the test-analysis samples were typically greater than 100, the adjusted reliability values are less than the original values.

The estimates of test reliability and scaled-score standard errors of measurement were available from test-analysis reports for each of the November and December forms of the SAT administered from 1970 to 1984. The reliabilities for a standard reference population were calculated for the study.

Trends. Changes in test lengths and testing time introduced in October 1974 could be expected to affect test reliability. The Spearman-Brown prophecy formula (see, for example, Gulliksen, 1987, p. 83) estimates the reliability of a test whose length is changed. Application of this formula shows that a reduction in SAT-V items from 90 to 85 items would lower reliabilities like those observed in 1970-73 by about .005, other things being equal. The formula also shows that a one-fifth reduction (from 75 minutes to 60 minutes) in the amount of testing time would have decreased SAT-V and -M reliabilities like those found in 1970-73 by about .02, had faster-moving items (Antonyms and Quantitative Comparisons) not replaced slower-moving items. The expected reduction in the reliability of the SAT-V should be kept in mind as one interprets reliability trends.

As Table 25 and Figures 8 and 9 indicate, throughout the 15-year period from 1970 to 1984, the SAT-V and the SAT-M tests demonstrated a high level of internal consistency. Except for the November 1978 SAT-M form, all of the SAT-V and SAT-M reliability coefficients were above .900. Despite some minor variation across administrations, no clear trends were evident in the reliabilities. The reliabilities of the December SAT-V forms fluctuated less than the reliabilities for the other SAT-V and SAT-M administrations. In 1974 the December SAT-V form was the only one of the four SAT-V and SAT-M forms that did not show a drop in reliability. The reliability levels were higher in succeeding years, indicating that shortening of the time limits of the test in the fall of 1974 did not have any lasting influence on the test reliability. In 1980 the reliability of the SAT-V test increased slightly (by .01) for the November administrations, and then remained at the higher level. For the SAT-M test forms administered from 1978 to 1984, the reliability coefficients either increased or remained stable. The reliability for the SAT-M appeared slightly elevated in November 1975, November and December 1976, and December 1984; the value for November 1978 was the lowest reliability observed.

Since the patterns of change in the adjusted reliabilities and the scaled-score standard errors of measurement are simply opposites of one another except for scale, only the former index is discussed here. The adjusted reliabilities fluctuated less than the actual reliabilities. Most of the adjusted reliabilities fell within the .90-.91 range for the SAT-V and .88-.89 for the

SAT-M. The patterns of change for the adjusted reliabilities showed little effect due to the shortening of the test. The adjusted reliabilities for the SAT-V for both November and December were relatively stable except for the lower value observed for the December 1973 form. There was a decrease in the November SAT-V adjusted reliabilities from 1982 to 1984. For the December SAT-V forms, however, the adjusted reliabilities between 1970 and 1973 were in general slightly lower than they were in succeeding years. The December SAT-V adjusted values exhibited a slight upward trend until 1981 and then a downward trend.

The adjusted reliabilities for the SAT-M were more variable than those for the SAT-V--particularly those for the December forms. No systematic trends were evident. For November the adjusted reliabilities were relatively stable, but showed a very slight downturn from 1981 to 1984. The adjusted reliabilities for the November 1978 and the December 1970 and 1974 forms stood out in the plots as low forms; the adjusted reliability for the December 1976 form stood out on the high side. For the December SAT-M forms, adjusted reliabilities fluctuated almost .03 between 1974 and 1976. Otherwise they were relatively stable until 1982, when they increased by about .01 to a level above that for the November forms. The increase in reliability noted for the December forms from 1978 to 1984 disappeared when the reliabilities were adjusted.

Test-Retest Reliability

The internal-consistency reliabilities indicated the interrelatedness of the items on the SAT. An alternate measure of the reliability of the test is the consistency with which a test yields similar scores for an individual across test administrations. Test-retest correlations for the SAT are based on students tested in the spring of the junior year on one form of the test and in the fall of the senior year on another form of the test. Alternate forms of the SAT are constructed to be as parallel as possible to one another in terms of difficulty and content. If the changes in the test introduced in 1974 affected reliability, one would expect to find lower correlations from the spring of 1973 to the fall of 1974 than in other years. If only the change in SAT-V test length affected reliability, then for the SAT-V one would expect a lower correlation in 1974 than in the preceding three years and even lower correlations after 1974 because of the reduction in test length from 90 to 85 items.

Table 26 and Figures 8 and 9 give the test-retest correlations for November and December forms administered from 1970 to 1984. The table provides data for a number of junior-year to senior-year testing patterns, whereas the graph includes only data from the within-year patterns with the largest sample sizes. Overall, the correlations remained relatively stable between 1970 and 1984, ranging from .87 to .89 for most values. For the SAT-V the March/April-to-November pattern reached a low of .88 from 1974 to 1978 before returning to the .89 level. Similarly, the March/April-to-December pattern for the SAT-V exhibited a low correlation of .87 from 1973 to 1980. The introduction of the 85-item SAT-V in October 1974 could have contributed to this slight downturn. However, all but one of the correlations from 1979 on were at the .89 level. The correlations in the later years were in general similar to those in the earlier years--no obvious trends are discernible.

The test-retest correlations for the SAT-M varied from .86 to .90 but showed no obvious trends. The average value was .88.

In general, the form-to-form variation in internal consistency reliabilities and test-retest reliabilities was small. The intern·l consistency reliabilities for the SAT-V and the SAT-M were all above .90, and the test-retest correlations fell into the .88 to .89 range for the most part. There was no evidence of any deterioration in test reliability over time. If anything, the more recent November and December SAT-V and SAT-M forms had reliabilities at least as high as those administered in the early 1970's.

Changes in Correlational Patterns

Changes in the SAT, if they had an effect, would presumably have affected the correlations with other variables as well as the statistics associated with the test itself. Trends in correlations of the SAT-V and the SAT-M with other variables and with each other and trends in the correlations of SAT sections and subscores provide evidence of the stability with which the test measures what it measures. If the test measured the same thing to the same degree of precision over time, these correlations would tend to remain relatively constant across years. Changes in either of the variables being correlated could, of course, result in instable correlations over time. Such correlations are affected by changes in the test-taking population as well. Nevertheless, an examination of the correlations among the SAT and other tests and among SAT sections and subscores should provide some evidence as to whether any changes introduced into the SAT from March 1970 to January 1975 were important enough to affect these correlations.

The correlations among the SAT-V, the SAT-M, and the Test of Standard Written English (TSWE), which was introduced in 1974, were available from test-analysis reports. Also available from these reports were correlations of scores from the two verbal sections and from the two mathematical sections for November and December SAT test forms. In addition, the test-analysis reports gave the correlations of the verbal subscores, Reading and Vocabulary. The only correlations available for the entire 15-year period were the correlations of SAT-V and SAT-M, and the correlations of the two verbal and the two mathematical sections.

Correlations Among the SAT-V, the SAT-M, and the TSWE

The various correlations among these variables are provided in Table 27 and Figure 10. The original correlations as well as those corrected for attenuation are given; the corrected correlations take account of differences in test reliabilities and indicate the extent to which the tests measure the same underlying construct. The correlations between SAT-V and SAT-M scores fluctuated somewhat for November and December SAT forms from 1970 to 1984, ranging between .62 and .71. The pattern for the correlations corrected for attenuation was almost identical to that of the original correlations, and clearly showed that the SAT-V and SAT-M were measuring different underlying constructs. Ignoring the outliers, there is a trend downward in the November uncorrected correlations from .68 in 1970-72, to .67 in 1974-81, to .66 in 1982-84. The November correlations from 1973 to 1978 and from 1981 to 1984 were relatively stable. The large decrease in 1979 followed by a large increase in 1980 was unusual and difficult to explain. In contrast to the November pattern, the December pattern showed more fluctuation. There was a large decrease in the correlation in 1974, but several high correlations occurred later on. There appeared to be a slight downward trend in the December correlations of about the same magnitude as that for the November correlations.

As expected, the correlations between the SAT-V and the TSWE were higher than were those between the SAT-M and TSWE and between the SAT-V and the SAT-M, since the SAT-V and the TSWE both measure verbal skills and abilities. On average, the correlation between the SAT-V and TSWE was .78-.79; the November correlations were slightly lower than the December correlations. The November 1974 correlation of .75 between the SAT-V and TSWE was the lowest observed for the November and December forms. For November there appeared to be a slight increase in the correlations between the SAT-V and the TSWE from 1974 to 1976 and then a leveling off. For

December the SAT-V and TSWE correlations reached a high point in 1978 of .81 and then dropped down to previous levels (.78 - .79) after that.

In general, the November and December correlations of SAT-M and TSWE scores ranged between .62 and .64. The November and December 1974 correlations (.59 for both), along with the December 1981 correlation (.55), were low points for the 15-year period. The November correlations increased from .59 to .63 from 1974 to 1976 and, after a slight decrease, reached a high value of .64 in 1981. December correlations exhibited an increase of .06 from 1974 to 1977 to a high of .65 and then a gradual decrease of .05. There was a decrease of .07 in the correlation in 1981, to a low of .55, followed by an increase of .10 in 1982. This decrease was apparently due to the December SAT-M form, which also had a low correlation with the SAT-V.

Correlations of Sections 1 and 2

The correlations of scores on the separately timed verbal or mathematical sections (see Table 28 and Figure 11) provide additional information relevant to whether changes in the content of the SAT affected the homogeneity of the test. These correlations measure the consistency of scores on two parts of the SAT. The sections contained somewhat different content (see Figures 1 and 2) and were not strictly parallel in difficulty. The two verbal and two mathematical sections differed in numbers of items, numbers of items of each item type, and testing time limits. If changes in the test introduced in October 1974 and afterward affected what and how well the test measured, one would expect slightly lower SAT-V section correlations from 1974 on relative to the correlations before 1974. Of particular interest are the correlations corrected for attenuation, which show whether the sections measured essentially the same underlying construct(s). The corrected correlations are the estimated correlations of perfectly reliable section scores.

The section scores correlated highly after correction for attenuation. The corrected correlations ranged from .97 to 1.00 for both the SAT-V and the SAT-M. Corrected correlations that are this high indicate that the SAT-V sections and the SAT-M sections, regardless of content and format changes, measured essentially one underlying factor from 1970 to 1984. In general, the mathematical sections had higher corrected correlations than the verbal sections--despite the fact that in each of the SAT-M forms, one of the mathematical sections always contained only Regular Mathematics items. There is a suggestion in the data for both November and December that the mathematical sections became more homogeneous after the introduction of the shortened SAT in October 1974. No such pattern emerged for the SAT-V. These data provided no evidence of any effect due to the changes introduced into the content of the SAT from 1970 to 1984.

The uncorrected section correlations for the SAT-V varied somewhat between November and December administrations. November correlations tended to be higher. There were no clear trends in the SAT-V data over time. The SAT-M uncorrected correlations for November and December were more similar than those for the SAT-V. The November correlations peaked in 1975 and 1976 and the December correlations, in 1976 and 1977 and then again in 1984. The correlations for the forms administered from 1974 on tended to be higher than those for the forms administered from 1970 to 1973. Like the SAT-V data, these data do not indicate an effect due to the restructuring of the SAT in 1974.

Correlations of Reading and Vocabulary

The correlations between the SAT-V subscores provide additional evidence of the essential unidimensionality of the SAT-V from 1974 to 1984 (see Table 29). For November and December forms from 1974 to 1984, the Reading and Vocabulary subscores correlated highly with one another. The uncorrected correlations were more variable than the corrected correlations and varied arcund a central tendency of .80. The correlations between these subtests, corrected for attenuation, ranged from .92 to .96, indicating that subscores were measuring essentially the same underlying construct. For much of the 11-year period from 1974 to 1984, the correlation between the Reading and Vocabulary subscores varied only slightly.

Summary Analysis of Changes in Statistical Characteristics

Thus far this chapter has focused on changes in a number of statistical characteristics of the SAT from 1970 to 1985 based on data from November and December SAT forms. Interpretation of these changes and the identification of trends is difficult because the size of the change is relative to the magnitude and variability of each particular characteristic, or variable. For instance, a change in the mean equated delta of .10 is relatively small; while a corresponding change in reliability is relatively large.

To help identify trends in the data, the data were subjected to a regression analysis to see whether an overall linear trend appeared across the 15-year period. This regression analysis produced a slope for each variable studied. The slope for a given variable was then standardized by dividing it by the standard deviation of that variable across November and December forms and across years. (See the numbers in the column headed "Overall Slope" in Tables 30a, 30b, and 30c.) These standardized slopes were thus comparable from one variable to another. This analysis identified drifts in the statistical characteristics of the SAT across the entire 15-year period. For instance, it indicated whether the test become progressively more or less speeded, or retained the same level of [un]speededness from 1970 to 1984.

From 1970 to 1984 the major change in the statistical specifications of the SAT occurred in the fall of 1974, when the time limits of each part of the SAT was reduced from 75 to 60 minutes. (See Table 12 for the statistical specifications for the SAT.) The planned changes affected the equated delta and to a limited extent the mean item-test biserial correlation. Differences in the statistical characteristics due to planned changes in statistical specifications would presumably appear as differences in the means for the years covered by the planned changes. To determine whether the actual means differed in the periods before and after the 1974 change in specifications, the standardized difference between the mean for the 1970-73 period and the mean for the 1974-84 period was calculated. (See the numbers in the column headed "Difference Between Period Means" in Tables 30a, 30b, and 30c.) Although the modifications in statistical specifications applied only to the equated deltas and the item-test biserial correlations, the difference between the period means was determined for each of the statistical characteristics studied.

While changes coinciding with planned changes in the statistical specifications would tend to appear as abrupt changes, differences in the other statistical characteristics of the test might appear as gradual trends in the data within each period. To identify possible trends within the 1970-73 and 1974-84 periods, within each period the data were analyzed to determine whether a linear trend existed. As with the previous regression analysis, the slopes were standardized by dividing them by the standard deviations of the variables. (See the slopes in the columns of Tables 30 to 32 headed "Slopes Within Periods.") Of interest was whether systematic linear increases or decreases occurred within each period.

These analyses were performed on combined November and December data. On occasion real differences in the November and December trends may have been present. The readers interested in such differences are invited to examine the tables and graphs that show the data for both months.

Test Difficulty

SAT-V. As with the SAT-M, the specified mean equated delta for each assembled SAT-V form was reduced from 11.7 to 11.4 in 1974. At this time the standard deviation of equated deltas was increased from 2.9 to 3.3. According to the trend analysis, the actual standardized mean and standard deviations of equated deltas were consistent with these changes in test specifications. (See the difference between period means in Table 30a.) The equated delta was on average lower for November and December forms administered from 1974 on than for those administered prior to the change in specifications. In addition, the standard deviation was on average higher for forms administered from 1974 on. Another measure of test difficulty, the scaled score conversion for the raw score midpoint, was also lower on average following the change in specifications. Therefore, the SAT-V was slightly less difficult after 1974 than it was from 1970 to 1973.

In 1982 the specified standard deviation of equated deltas was changed from 3.3 to 3.0. This change presumably was reflected in the slightly negative standardized slope for the standard deviations for the November and December SAT-V forms. This slope, however, was small relative to the slope of .50 for the 1970-73 period, during which the standard deviation should have remained constant.

SAT-M. In 1974 the specified mean equated delta for each assembled SAT-M form was decreased from 12.5 to 12.22. Consistent with this change, the trend analysis indicated that the mean equated delta was lower for forms administered from 1974 on than for those administered prior to the change in specifications. There was also a substantial increase in the average standard deviation of equated delta. Additionally, the scaled score which corresponds to the raw score midpoint was lower during the period subsequent to the change in specifications. Therefore, the SAT-M was slightly less difficult after 1974 than it was from 1970 to 1973.

While there does not appear to be any noticeable trend across the entire time period, between 1970 and 1973 the means and standard deviations of equated deltas increased somewhat, suggesting a more difficult test. During this period the scaled score corresponding to the raw-score midpoint had a negative slope, providing evidence of a decrease in test difficulty. A possible shift in the delta scale was suggested earlier in the report as an explanation of this inconsistency.

Item-Total Test Correlations

SAT-V. In 1974, the specified mean item-total test biserial correlation was increased from .42 to .43. The difference in standardized means clearly reflects this shift. In 1982 the item-test correlations were constrained to be between .41 and .45. The slightly positive slope in the 1974-84 period provides only scant evidence that the mean biserial correlations were higher from 1982 on.

SAT-M. The specified mean biserial correlation was .47 from 1970 to 1984. Therefore, no trends were expected. Nevertheless there an increase in the mean biserial correlation was evident between the 1970-73 period and the 1974-84 period. Within the period prior to the shortening of the SAT, there was a slight increase in the item-test correlations.

Relative Test Difficulty

SAT-V. Patterns exhibited by the mean observed delta and the mean adjusted proportion correct, the two measures of the relative difficulty of the test, were consistent for the SAT-V forms. The SAT-V forms tended to be slightly less difficult for the examinees taking the test from 1974 to 1984 than for those taking the test from 1970 to 1973. The standard deviation of observed delta exhibited a relatively large increase between the two periods, indicating more varied test performance among later test takers.

SAT-M. Although the absolute difficulty of the SAT-M changed in 1974, when the specified mean delta was decreased, the relative difficulty of the test did not appear to change across the 15-year period of this study.
The mean adjusted proportion correct, however, did show a very slight decrease from the 1970-73 period to the 1974-84 period, indicating a slightly harder test for the group taking it. A decrease in the standard deviation of observed delta was evident within the 1970-73 period and between this period and the 1974-84 period. The decrease suggests either a more homogeneous test or a more homogeneous group of test takers.

Speededness

SAT-V. While no clear linear trends in speededness exist across the 15-year period or within periods, both sections of the SAT-V, but especially Section 1, the longer section, were clearly more speeded after the shortened SAT was introduced in 1974.

SAT-M. The trend analysis of the speededness indices demonstrated that Section 1 of the SAT-M was considerably less speeded from 1974 to 1984 than it was prior to this date. The percentage of test takers completing 75% of the test increased, while the mean number of not-reached items and the ratio of not-reached and total-test variance decreased. Throughout the time period of this study, Section 2 of the SAT-M remained a rather unspeeded section. Although the mean number of not-reached items increased after 1974, the other speededness indices displayed little or no change from 1970 to 1984.

Reliability

SAT-V. The internal-consistency reliabilities showed little change across the 15-year period. On the other hand, the scaled-score standard error of measurement decreased somewhat between the two periods, and correspondingly the adjusted internal-consistency reliability increased. Thus, the test in later years measured at least as well as the test in the earlier years. Test-retest correlations for March/April test takers who repeated the test in November or December decreased somewhat during the 1970-73 period. (These patterns tended to have the largest repeater volumes in November and December, respectively). These test-retest correlations were also slightly lower after 1974.

SAT-M. During the 15 years reviewed in this study, the internal-consistency reliability coefficients for the November and December SAT-M forms were generally relatively high, at or above .90. Nevertheless, a trend toward higher reliabilities was evident in 1970-73 and between this period and the 1974-84 period. When these reliabilities were adjusted to have a standard deviation of 100, reducing the values slightly, the upward trend in 1970-73 was still apparent. No change was evident in the test-retest correlations for the spring (March/April) to fall (November or December) patterns. See Table 26 for data on other repeater patterns.

Correlational Patterns

Only a few meaningful patterns appeared in the correlational data. One pattern of note was the lower correlations for SAT-V with SAT-M from 1974 on. On the other hand, the correlations between sections 1 and 2 of the SAT-M were much higher in the later period. Thus, it appears that the factors measured by the SAT-M and the SAT-V were less highly related after 1974, and that the SAT-M was a more heterogeneous test. The only other relatively large changes were two trends in 1970-73: a decrease in the correlations between the two SAT-V sections and an increase in the correlations between the two SAT-M section. Given the few changes, the overall picture is one of relative stability in the SAT correlational patterns.

Changes in Statistical Characteristics and Trends in Predictive Validity

Presumably, changes in the test, if they were important enough to influence predictive validity, would also have influenced the statistical characteristics of the test. Changes in test length and test difficulty, for example, would likely affect the reliability of the test scores if they had any effect on predictive validity. If predictive validity depended on changes in the test, one would expect to see similar patterns for SAT predictive validity and test-related statistics. This section of the report discusses the relationship of changes in test difficulty, speededness, reliability, and correlational patterns to trends in predictive validity. Table 31 summarizes changes in the statistical specifications. The previous section reported the trends in the various statistical characteristics of November and December forms given from 1970 to 1984.

Test Difficulty

Changes in test difficulty could affect the validity of a test if the changes caused the test to lose measurement power in the score ranges of importance to the colleges conducting validity studies. When the SAT was shortened in 1974, the SAT-V and SAT-M mean equated item difficulties (equated deltas) were each lowered by .3, thus making the tests easier. The distributions of item difficulties were also changed--reducing the number of items of middle difficulty while providing a larger number of more difficult items (see Table 12). Then in 1982, the SAT-V item difficulty distribution was changed again--this time to decrease the dependence upon difficult items. The section on test difficulty noted that the test deviated from specifications somewhat and if anything the test was slightly less difficult than intended. In general, the intended changes in difficulty were satisfied.

Test difficulty and reliability. Changes in test difficulty are intimately related to changes in test reliability, provided that what the test measures is consistent over time. The new item-difficulty specifications for the shortened SAT, derived from test-design analyses using item response theory, were intended to preserve measurement power in the middle-to-upper part of the score scale, yet make the test more appropriate for the average test taker. The analyses estimated that, relative to a 90-item form of the SAT-V administered in 1971, the 60-minute, 85-item SAT-V would provide considerably more measurement power below 300 and less measurement power throughout the rest of the scale (Braswell & Marco, 1974). Similar estimates for the 60-minute, 60-item SAT-M showed superior measurement below 450 and above 650 relative to a 1970 form but a loss of measurement power in the 450-650 score region. The estimated loss of measurement power occurred in the neighborhood of the SAT-V and -M means for a typical college validity sample (443 to 480 for the SAT-V and 479 to 509 for the SAT-M--see Tables 2 and 3).

Theoretically, changes like these in difficulty could have affected the reliability of the test for validity-study samples consisting primarily of students who scored in the middle part of the score scale, and consequently the predictive validity for these groups. Predictive validities for selective colleges and colleges with low-scoring students would presumably have been less affected. However, reliability estimates using item-response-theory procedures prior to the time the changes were introduced showed negligible effect on the reliability coefficient, an index of the overall measurement power of a test for a group of test takers. The analyses estimated a .01 decrease in the reliability of the SAT-V and a .002 decrease for the SAT-M as a result of reducing the number of SAT-V items and reducing the difficulty of the SAT-V and the SAT-M. Such a small loss in reliability would not cause any deterioration in predictive validity.

The other change in specifications during the 1970-85 period occurred in January 1982, when the specified distribution of item difficulties for SAT-V was changed (see Table 12) because of the scarcity of difficult items. The specified number of items in the delta 13-14 range was increased from 16 to 24 items, and the specified number of items at delta 15 and above was decreased from 16 to 8. The new specifications called for a distribution of item difficulties much like that used prior to 1966. Analyses based on item response theory estimated an improvement in measurement power, relative to forms administered in 1977 and 1979, in the 350-600 score range and a loss of below 350 and above 600 (Petersen, 1981). This change would presumably have improved the reliability--and therefore the validity--of the SAT for the typical college validity-study sample relative to the forms used between 1974 and 1981.

Trends in the data. The effect of test difficulty on reliability and validity depends on the extent to which the test met difficulty specifications and the ability of the students in the colleges who chose to conduct validity studies over time. Tables 13 and 14 show that the means and standard deviations of item difficulties for November and December forms were in general close to their specified values. If there was a trend, it was for more recent forms to be slightly easier than specified. The distributions of item difficulties for the November and December forms deviated somewhat from specifications but averaged across several forms were consistent with specifications (see Tables 15 and 16 and Tables B-7 to B-10 in Appendix B). The data on score conversions (see Tables 17, 18, and 19), however, indicated that the SAT-V and -M became easier even after the easier SAT was introduced in 1974.

Tables 2 and 3 show that college students represented in the validity-study samples who took the shortened SAT (entering classes from 1975 to 1985) were in general less able than those that took the longer forms. The SAT-V means ranged from 443 to 480; and the SAT-M means, from 479 to 509. After a fairly substantial drop of about 15 points in the SAT-V and -M means in 1977, the means remained relatively stable from 1977 to 1984. Despite these changes, the means were still in the region where a loss of measurement power was expected as a result of the change in SAT specifications. Thus, it is theoretically possible that changes in difficulty caused a loss of reliability, which in turn resulted in lowered validity. The section on reliability addresses this issue directly.

Relative Test Difficulty

Relative test difficulty reflects both actual test difficulty and the ability of the test takers and thus is an indicator of how well the test measures a group of individuals. If test difficulty decreases in proportion to decreases in test-taker ability, relative test difficulty will not be affected. Since the ability level of college-bound seniors as well as validity-study samples declined from 1971 to 1980 before showing an upturn, relative test difficulty provides a better indication of the measurement power of the test for these groups than does test difficulty per se.

To provide the most measurement power for a group of individuals, test should theoretically be halfway between chance success and the maximum score (middle difficulty). In the case of the SAT, this value equals .50 in terms of the mean adjusted proportion correct, or alternatively 12.0 (for the SAT-V) or 11.9 (for the SAT-M) in terms of the mean observed delta (see Educational Testing Service, 1989). Theoretically, item difficulties should also be peaked at middle difficulty. (Of course, the SAT is intended to measure relatively well throughout the score scale and not just at the score level of the average test taker and thus uses a wide range of item difficulties.) The previous discussion has already noted that with the change in the specifications in 1974 came a loss of measurement power from 300 to 800 on the SAT-V and from 450 to 650 on the SAT-M. Some measure power was recouped for the SAT-V in the 350-600 range when the specified distribution of item difficulties was changed in January 1982.

The data in Table 20 and Figure 5 in fact show a decrease in relative test difficulty for November test takers from 1970 to 1984. The SAT-V mean adjusted proportion correct reached a high of .44 in 1984. The trend in mean adjusted proportions correct for December forms was uneven but also reached a high point of .39 in 1984. The SAT-M December mean proportions correct also peaked in 1984 at .40, and the November mean in 1983 at .45. Since validity-stu'y samples tended on average to score 10 to 20 points higher than November test-analysis samples, their mean proportions correct on these forms, if computed, would have been even higher--although still less than .50. These data suggest that the test was at least as appropriate for test takers in the later years as in the 1970-73 period. No decline in validity would be expected.

It is important to understand why the SAT tended to be easier for the test-analysis sample in the latter part of the 1970-84 period--despite a fairly dramatic decline in the average ability of test takers, particularly in December, and stable test-difficulty specifications from 1974 to 1984. The answer is found in Table 19 and Figure 3, which give the scaled scores corresponding to raw-score midpoints. Even though there was no intended decrease in test difficulty, the data clearly show that test difficulty eased during the 1970-84 period. Thus, the measurement power of the test was protected, albeit inadvertently, because the test became easier.

In saying this, one must at the same time call attention to the difference between score conversions, which resulted from score equating, and mean equated deltas, which resulted from delta equating. The latter did not show as much of a decrease in test difficulty from 1974 to 1984. Because of the rigor of the score-equating process (see the next part of the report), however, equating results are more trustworthy than delta equating results. Hence, the conclusions stated here are based on the score conversions.

Speededness

Relation of speededness to reliability and validity. The speededness of a test also could affect its measurement power. In the strictest sense, almost all timed tests are somewhat speeded in that not all test takers are able to reach all items in the tests. However, when nearly all test takers are able to attempt most questions, the test may be considered unspeeded from a practical point of view. The extent to which the speededness of a test might affect predictive validity depends on the relationship of rate of responding to the construct measured by the test. Within limits, if the rate of answering is highly related to the knowledge, skill, or ability being measured, increased speededness will increase the predictive validity of the test. The SAT is a test of developed verbal and mathematical abilities. Presumably, the more highly developed verbal and mathematical abilities are, the faster the rate of response is to test items and the more accurate the item responses are. That is, increased ability affects both accuracy and rate of response.

For a test of developed abilities, a test that is slightly speeded for the test taker probably measures better than a test that is unspeeded. If the testing time for the SAT were adjusted to make it less speeded, more test takers would finish the test in the time allotted, thus decreasing ability of the test to distinguish among high-ability test takers but increasing the measurement power of the test among those of lower ability. On the other hand, if the test were made more speeded, the measurement power of the test would probably increase for the very highest scorers but decrease for the low scorers. The time limits for the SAT are set to enable it to measure well in the middle-to-upper parts of the score range.

It should be clear that making the SAT more or less speeded could increase or decrease the predictive validity of a test, depending upon the ability of the group. In general, however, reduced speededness would be expected to increase the measurement power of the test for the average test taker. The first question is, Did speededness change from 1970 to 1974? If it did not, it could not have affected validity. If it did, were changes in speededness consistent with the trends observed in predictive validity?

Summary of trends in speededness. As Table 21 to 24 and Figures 6 and 7 show, throughout the time period the percentage of test takers completing most items in the sections was typically over 95%, and the ratio of the not reached variance to the section-score variance was considerably less than .25, indicating that the SAT-V and the SAT-M were not very speeded for the average test taker. The speededness indices show that, except for the November 1974 form, the shorter section of the SAT-V was relatively unspeeded. The longer SAT-V section was relatively unspeeded until 1976 for the November form and until 1977 for the December form, and since then was more speeded than the other section for all but the December 1984 form.

Both SAT-M sections tended to be relatively unspeeded for November test takers during the entire 15-year period--even more so since the introduction of the shortened SAT in October 1974. Section 1 of the SAT-M, the 25-item Regular Mathematics section, tended to be less speeded after the introduction of the shortened SAT than it was before in both November and December. This section was slightly more speeded for December test takers. Section 2, on the other hand, became more speeded in 1974, presumably because of the introduction of Quantitative Compaıison items or their placement in the section. In the following year the section was no more speeded than it was before 1974 and tended to have about the same degree of speededness as Section 1. Apparently, the placement of the Quantitative Comparison items in the middle of the section succeeded in reducing speededness.

The tendency of Section 2 of the SAT-V to become more speeded might have caused the SAT-V to measure average test takers less well and thus decreased the measurement power of the test. On the other hand, the decrease in speededness for the SAT-M would presumably have increased the measurement power of the test for the average test taker. Since neither test was very speeded, one would expect small, if any, effects on reliability (measurement power) and validity.

Reliability

Expected changes in reliability. Reducing the administration time for each part of the SAT from 75 minutes to 60 minutes in the fall of 1974 would have resulted in a 72-item SAT-V and a 48-item SAT-M if test developers had constructed tests parallel in content to the previous versions. The Spearman-Brown formula,

$$r_{x'x'} = \frac{kr_{xx}}{1 + (k - 1)\ r_{xx}},$$

gives the reliability of a test k times as long as the original test (see Gulliksen, 1987, p. 78). Applying this formula to a test reduced by 1/5th shows that this shortening would have reduced reliability of either test by about .02. Because test developers increased the proportion of Antonyms in the SAT-V and introduced Quantitative Comparisons in the SAT-M--items that could be answered at a faster rate--the shortened test contained the same number of SAT-M items (60) and almost the same number of SAT-V items (85 *vs.* 90). The "faster" items were used expressly to ensure that the reliabilities for new forms of the test would remain at about the same levels as the those for previous forms. Application of the Spearman-Brown formula shows that a reduction from 90 to 85 items would reduce the reliability of the SAT-V, usually in the neighborhood of .91 to .92, by about .005. A test of the same length as the original test, as in the case of the SAT-M, would of course not suffer any reduction in reliability.

Expected changes in validity. The formula from classical test theory that relates changes in validity to changes in reliability is

$$r_{x'y} = r_{xy}\ \frac{\sqrt{r_{x'x'}}}{\sqrt{r_{xx}}},$$

where r_{xy} is the correlation of criterion Y and predictor X', which is equal to predictor X with a reliability of $r_{X'X'}$ instead of r_{xx} (Gulliksen, 1987, p. 94). Application of this formula demonstrates how shortening the SAT would affect predictive validities. For example a test reliability of .922, the reliability of the November 1973 form of SAT-V, would decrease to .918 for a test shortened from 90 to 85 items. A test validity of .41, the validity coefficient for the 1974 entering class, would stay at .41 as the result of this change in reliability.

Table 32 shows the expected effect of shortening the test in 1974 on predictive validity for several initial values of reliability and validity. The Spearman-Brown formula from classical test theory was used to estimate the reliability of a shortened test. This value was substituted in the formula for estimating the validity of a test whose reliability changed from r_{xx} to $r_{X'X'}$. The table shows the results of shortening a test by 1/18th, as in the case of the SAT-V, the length of which was shortened from 90 to 85 items in 1974. It also shows the results of shortening a test by 1/5th, as would have been the case had the SAT-V or the SAT-M been shortened from 75 to 60 minutes without altering the mix of items in the test. The original reliabilities and validities listed in the table are intended to cover the range of reliabilities and validities observed for the SAT-V and the SAT-M from 1971 to 1985.

One can see that shortening the test by as much as 1/5th has only a slight effect on reliability and has little effect on test validity when validity is in the neighborhood of .35 or .40. Even when reliability decreases by .03, from .84 to

.81, validity decreases by only .01.

Another way of looking at the possible effects of reliability on validity trends is to estimate how large a change in reliability would have been needed to account for the validity changes. Solving the equation that adjusts validity for changes in reliability for $r_{x'x'}$ yields

$$r_{x'x'} = \frac{r_{x'y}{}^2 \, r_{xx}}{r_{xy}{}^2} .$$

Applying this formula shows that changes in reliability could not possibly have caused the changes in SAT validity.

To increase validity from .36 to .41, the increase in the observed validity of the SAT-V from 1971 to 1974, would require an increase in reliability from .916, the reliability of the November 1970 form of the SAT-V, to an impossible value of 1.19. The corresponding value for the SAT-M, whose observed validity increased from .32 to .41 in the 1971-74 period, is 1.48! Similarly, to decrease validity from .41 to .32, the decline in SAT-V validity from 1974 to 1985, would require a decrease in reliability from .922, the reliability of the November 1973 form of SAT-V, to .562. This is the reliability of an SAT shortened from 90 to about 10 items. The corresponding value for SAT-M, whose validity declined from .41 to .31, is .522. An SAT-M consisting of about 6 items would provide su:h a reliability.

Summary of trends in reliability. The reliability coefficients shown in Table 25 and Figures 8 and 9 confirm that SAT test reliabilities for the shortened SAT did not suffer any noticeable loss in reliability. Although there was a decrease in reliability for all but the December SAT-V form in the initial year, the reliabilities were higher in succeeding years. The SAT-M internal-consistency reliabilities tended to increase or remain stable.

Test-retest correlations, shown in Table 26, in the transition periods compared with those for the other administrations provide additional general evidence about SAT reliability. If the SAT were affected when the testing program introduced the shortened SAT, or IRT equating, or some other change, then one would expect to see lower test-retest correlations from spring of the junior year to fall of the senior year in the transition periods than in the other periods. SAT test-retest correlations were relatively stable for November and December forms administered from 1970 to 1984, ranging from .87 to .89. The only sugg sstion of a decline in reliability came from the SAT-V March/April to November and December patterns. The November pattern showed a decrease of .01 from 1974 to 1978, and the December pattern a decrease of .02 from 1973 to 1976 and from 1978 to 1980. The correlations tended to increase slightly for the November administrations of the SAT-M. This trend, however, is very slight and of little practical importance. Little evidence is found for the trends in SAT predictive validity.

It is clear that changes in reliability did not account for the trends in validity. As Table 32 shows, slight variation in test reliability would have had a negligible effect on SAT-V or SAT-M predictive validity. Other factors must have been operating to account for validity trends from 1970 to 1985.

Correlational Patterns

For the predictive validity of a test to remain stable over time, the construct measured by the test and criterion measures should not change. Changes in the degree to which a test measures a construct would presumably affect predictive validity--provided that the predicted criterion did not change. Changes in the extent to which a test measures a construct would be expected to manifest themselves in changes in correlations with other tests and also in correlations among subsets of test items, such as section scores or subscores.

Examination of the internal structure of a test and the test's relation to external variables over time can provide information about the stability with which the test measured what it was supposed to measure. One would expect predictive validity to mirror changes in the test's internal and external relationships if the latter were in any way associated with trends in predictive validity.

Analyses of the reliabilities of November and December forms of the SAT from 1970 to 1984 indicated that, on average, the SAT-V and the SAT-M items consistently measured their respective constructs. These internal-consistency measures provided evidence about the degree to which each of the test forms was unidimensional. When the test is composed of more than one item type, however, the item types may differ in the degree to which each measures the same construct. These differences would tend to attenuate overall reliability of the test, but otherwise would go undetected.

Correlations Among the SAT-V, the SAT-M, and the TSWE. As noted earlier in this report, the observed correlations of the SAT-V with the SAT-M ranged between .62 and .71, indicating that the constructs measured by the two test are similar but not the same. Table 27 and Figure 10 show that the correlations of the SAT-V with the SAT-M tended to decrease from 1970 to 1984. The decrease was only about .02, however. The correlations fluctuated considerably, however, particularly those for the December forms. The small difference between the correlations before 1974 and those afterward suggests that the predictive validity of the SAT with freshman grades could have been attenuated slightly. On the other hand, a decline in the SAT-V and SAT-M correlations could have improved the validity of the SAT in that the two tests were measuring with less redundancy than before. Given the other evidence cited of the stability of the SAT-V and -M over time, the slight decrease in correlation is not likely to have had any effect on predictive validity.

The correlations of the SAT-V and the SAT-M with the TSWE, which was introduced in 1974, fluctuated to a minimal degree throughout the period from 1974 to 1984. Correlations of the SAT-V with TSWE were higher than were those between the SAT-M and either the SAT-V or the TSWE, since the constructs measured by the SAT-V and the TSWE were more similar to one another than either was to that of the SAT-M. The stability in these correlations over time suggests that there was no change in the constructs measured by the SAT that could have affected its predictive validity in this period, assuming no change in the criterion.

Correlations between sections 1 and 2. The correlations between the SAT-V and SAT-M sections corrected for attenuation indicated that the sections were measuring essentially the same underlying variable over time. The corrected coefficients ranged from .97 to 1.00. The SAT-M apparently became more homogeneous from 1974. However, no effect on trends in predictive validity would be expected.

Correlations between Reading and Vocabulary. The Reading and Vocabulary subscores were slightly less correlated than the SAT-V sections but still correlated highly. The correlations averaged about .94 after correction for attenuation. Again no obvious effect on predictive validity was likely.

Correlations among item types. Correlations among the SAT-V and SAT-M items types were discussed in the section on changes in the content of the test (see Table 10) and will not be repeated here. The correlations are, however, relevant to the issue of whether what the test measured stayed relatively constant over the period studied. Suffice it to say that those correlations did not indicate any major shift in what the test was measuring. For the most part, the SAT-V correlations among item types remained stable. Changes occurred primarily in the relationship between Reading Comprehension and other SAT-V item types. From the 1971 form to the later forms, the Reading Comprehension items were less correlated with the other verbal items in the forms administered. The correlations among item types also showed that replacing Data Sufficiency items with Quantitative Comparison items resulted in no discernible changes in the relationship between SAT-M and SAT-V item types. These data demonstrated that the SAT-V and SAT-M have remained essentially unidimensional tests. Therefore one would not expect changes in predictive validity as the result of changes in correlations among item types.

To conclude, from 1970 to 1984 the correlational patterns of the SAT-V and the SAT-M were reasonably stable. Although there were a few data points that deviated from the average, particularly for the December data, there was no evidence of any systematic trends in the data. The only correlations that suggested any change in the correlational patterns over time was the slight drop in the correlation between the SAT-V and the SAT-M, particularly for November test takers. This decrease was too small to have more than a marginal effect on predictive validity.

5. CHANGES RELATED TO EQUATING

Equating is the process by which scores from different forms of the SAT are placed on scale together. Equating is necessary because test forms vary somewhat in difficulty and other statistical characteristics despite being assembled to rigorous content and statistical specifications. Equating, which is conducted separately for the SAT-V and the SAT-M, occurs when a new form or revised form of the SAT is first administered. The result of equating is a table that converts raw scores on a test form to scaled scores, which are reported to test takers. Score conversions were discussed in the test difficulty section of the previous part of the report and therefore are not discussed in this section. This section first provides a description of SAT equating procedures in use frrom 1970 to 1985. Later sections provide information about trends in statistical indices associated with the various equatings. The last section relates these trends to trends in SAT predictive validity.

Changes in Equating Procedures

Equating Design

Since 1954, SAT equating has used the anchor-test equating design involving items in common with two old forms (Donlon & Livingston, 1984). The anchor test provides the data needed for taking account of differences in total test samples selected from different test administrations. The anchor test, which throughout the 1970-to-1985 period consisted of either 40 verbal items or 25 mathematical items, is usually administered in the 30-minute variable section of the SAT and TSWE test booklet. Thusly administered, it is an "external" anchor test--i.e., external to the operational test and thus not counted in the operational test score. In a few instances in the early 1970's, the anchor test was administered as part of the operational SAT and thus was "internal" to the operational test rather than external.

The anchor-test design, whether internal or external, was in use throughout the 1970-to-1985 period. Chapters in the Admissions Testing Program's technical handbooks (Donlon & Angoff, 1971; Donlon & Livingston, 1984) describe the anchor-test equating design and technical procedures used with the SAT in detail. The data for equating come from new- and old-form samples. A sample of test takers from a current administration takes the new form of the SAT and a verbal or mathematical equating test that was previously given with an old form of the SAT. Four equating tests provide data for a given equating: two link the new verbal form to two old verbal forms and two link the new mathematical form to two old mathematica forms.

Braiding System

Each new form of an SAT is usually linked to two old forms according to a braiding system that protects against the development of "strains" among the test forms. Without such a system it is likely that over time forms from a particular administration would develop their own scale--one that would deviate somewhat from the scale of other SAT forms. The braiding system specifies the two old forms to which a new form is to be linked. The general principle driving the braiding system is that over time forms from a given administration are linked to forms from each of the other administrations in a balanced way. The braiding system has changed to take account of the increase in the number of new forms administered in a testing year, but the objective has remained constant--to ensure the integrity of the score scale by protecting against the development of administration-specific strains.

Figure 12 gives the braiding system for the SAT-V forms introduced from March 1970 to January 1985. The braiding system for the SAT-M forms differed only in a few instances and is not shown. In the figure, each box identifies the primary form introduced at the Saturday administration in the month and year indicated by the column and row. At some administrations two or more forms were introduced, but only one was included in the braiding system for future equatings. Arrows indicate the two old forms to which the new form was equated. Boxes with no arrows leading to them denote forms that did not serve as parents for other forms. Some, such as those shown for 1971 to 1975, used old-form equating items as part of the operational test, with no direct provision for future equatings. Others, such as form 1 in January 1981, were dropped from the braiding system because of the difficulty of equating.

There were several exceptions to the typical procedure of equating to two previously administered old forms through two external anchors during the 1971 to 1978 time period. Sometimes an external equating section common to an old form was used as an operational section of a new form. Also, once in a while one of the two old forms was given at a current administration along with the new form. Moreover, when more than one new form was introduced at the same administration, they often used the same external anchor tests and were equated back to the same old forms. These exceptions caused some forms to be used as parent forms three or four times and others to be used only once. To prevent any long-term impact on equating, the braiding system was altered so that only one of the two new forms equated through the same set of common items to two old forms could be used as a parent form.

Another change occurred after the expansion from five to seven administrations per year in 1979. The same or similar old-form equating assignments were used several times, threatening the development of equating strains. For example, the December form was equated to a May form in 1980, 1981, 1983, and 1984. This particular problem was identified and corrected during the 1984-85 testing year. Although the braiding system had to be revised from time to time to avoid problems that developed, its use has protected the scale against strains and has helped ensure scale stability.[1]

Equating Methods

Various equating methods can use the anchor-test data collected for the SAT. Donlon and Angoff (1971), Angoff (1982), Marco, Petersen, and Stewart (1983), and Donlon and Livingston (1984) have provided descriptions of these methods. The methods used operationally with the SAT in the 1970 to 1985 period were the following: Tucker observed-score, Levine equally reliable, Levine unequally reliable (used when total tests were of different lengths), equipercentile through an anchor test, and item-response-theory (IRT) three-parameter-logistic true-score. Lord (1980), in his book on applications of IRT, gave a detailed account of the IRT true-score equating method used with the SAT. Petersen, Cook, and Stocking (1983) also provided a detailed description of this method. The last two methods produce curvilinear conversions; the other three, linear. The Levine and IRT methods are true-score methods in that they use estimated true-score relationships to derive score conversions. True-score equating is considered to give more accurate conversions when samples or test forms differ considerably in their characteristics.

Table 33 identifies the equating methods that provided the conversion lines for the November and December forms of the SAT administered from 1970

[1] The authors are grateful to David J. Wright for assisting with the preparation of the section on braiding.

to 1984. IRT equating was first used with the January 1982 SAT to take account of the revised item-difficulty specifications for the SAT-V. Levine equating was used whenever the new- and old-form samples differed considerably in ability level.

The SAT conversion line is an average of the two new-form-to-old-form equatings. The tables show that, except for the 1979 SAT-V line, all of the SAT-V and -M November lines from 1970 to 1981 were averages of the two Tucker lines. And except for the November 1984 SAT-M equating, all of the November lines from 1982 to 1984 were averages of the IRT lines. In contrast, the equating decisions for the December forms varied considerably. Almost half of the SAT-V and -M lines in the earlier period were Tucker-Levine decisions. Moreover, some of the conversion lines from 1982 to 1984 were not pure IRT conversions. The greater variety of equating lines for the December equatings was due to the larger ability differences between new- and old-form samples.

Changes in Equating Indices

Equating is most effective when the new- and old-form samples are as equivalent as possible on any and all characteristics. Single-group and random-group equating designs assure this equivalency. In SAT equating, however, new- and old-form samples come from different administrations and are nonequivalent to some degree. Statistical procedures such as those used in equating cannot compensate completely for differences between nonequivalent samples. If the equating test is very similar to the total test, as in the case of the SAT, and correlates highly with this test, then equating is more likely to provide appropriate statistical adjustments.

Ideally, equating lines that result from the two separate equating legs contributing to the operational equating line agree with one another. High agreement, or reliability, does not necessarily ensure high validity. That is, equating results could still be biased compared with the results from an ideal equating even if the two equating lines agreed. Moreover, an unbiased equating line could result from averaging two lines that disagreed Nevertheless, high reliability is desirable. Statisticians tend to distrust more the score conversions that come from lines that disagree.

Several indices can be used to evaluate equatings. These were derived from the data on equating samples, which were available from the SAT testing program statistical files.

Equating Samples

Since most new forms of the SAT have been equated back to two old forms, data are usually available for two new-form samples and two old-form samples. Like the test-analysis samples, the equating samples were selected from the total population before 1981. Then, because the number of younger test takers increased, equating samples were restricted to high-school juniors and seniors, the group for which the test is designed. The new- and old-form equating samples for the November and December adminisrations from 1970 to 1984 furnished data for the current study. Tables 34 and 35 provide descriptive statistics on these equating samples. Their means were similar to those for the November and December senior test takers (see Tables 2 and 3).

Equating indices

Individual indices. There are a number of indices that can be used to evaluate equating. The indices available for use in this study were (1) differences in equating-test means and variances between the new-form and old-form equating samples, (2) the difference between the two equating lines on which the operational conversions were based, and (3) correlations of equating-test and total-test scores.

Two indices were used to measure ability differences:

$$\frac{|M_{new} - M_{old}|}{S_{new+old}},$$

the absolute standardized difference between equating-test means, and

$$\frac{S_{new}^2}{S_{old}^2},$$

the ratio of equating-test variances, where M_{new} and S_{new}^2 are the equating-test mean and variance, respectively, for the new-form sample, M_{old} and S_{old}^2 are the equating-test mean and variance, respectively, for the old-form sample, and $S_{new+old}$ is the equating test standard deviation for the combined new- and old-form samples. These indices are calculated routinely in equating and are used to help decide whether to use Tucker or Levine equating. Levine equating is generally preferred whenever the absolute value of the difference in means is greater than or equal to .25, or the ratio of the variances is less than or equal to .8 or greater than or equal to 1.25. These two indices were available from equating files for each pair of new-form and old-form equating samples used in a particular equating.

The measure of how well the equating test paralleled the total test, was the correlation between the scores on these tests for each of the new- and old-form samples. Since a few of these correlations were based on equating tests that were a part of the total test (internal equating tests), these particular correlations were adjusted to make them comparable to correlations based on external equating tests. Procedures developed by Angoff (1956) were used for computing the adjustment. (One of the SAT-V correlations for November 1974, however, could not be adjusted because it was based on a total test that included some but not all of the equating-test items). A trend of lower correlations over time would not only result in a larger standard error of equating but would also suggest less adequate correction of ability differences between samples.

The other index, used to measure the difference between the two equating lines on which the operational conversions were based, was simply the absolute value of the difference between the scaled scores produced by these two equating lines at the midpoint of the raw-score range. This difference ignores differences across the full score range, but captures that part of the score range where the data are relatively dense.

Composite index. Because it is difficult to interpret information on so many indices, individual indices for a given administration were combined to form a composite index of equating. The composite index should not be interpreted as a measure of equating quality, for it is based only on the individual indices, which include only a few of the possible indices that could be used to measure equating quality. Also, because the scale for the composite index is based on the distributions of the individual indices, and thus is normative, a low average composite index should not be interpreted as low in an absolute sense. The equating methods used for the SAT ensure that SAT equatings are of high quality. A low composite index indicates only that the index was low relative to the indices for the other equatings.

The composite index is admittedly arbitrary, but is based on some logical considerations. The overall composite was computed by weighting each of the two equatings and the difference between the two equating lines equally. For the first and second equatings, individually, the absolute standardized difference between means on the equating test was weighted twice as much as the ratio of the equating-test variances. These variables measure the ability differences between samples, and are used operationally to choose between true-score and observed-score equating. The difference between equating lines is considered important in evaluating operational equatings. Differences that are large (say 20 points or more) are considered problematic and may lead to weighting one of the lines less than the other line. The ratio of the variances and the two equating-test-total-test correlations were given equal weight in the composite index. The correlations, while also important, are generally high and therefore tend to have less effect on equating results.

The weights assigned to the individual indices were as follows:

1. First Equating or Second Equating (total of 8 points each)	
a. Standardized Difference Between Equating-Test Means:	4
b. Ratio of Equating-Test Variances:	2
c. New Sample Equating-Test-Total-Test Correlation:	1
d. Old Sample Equating-Test-Total-Test Correlation:	1
2. Difference Between Equating Lines:	8

In the one instance in which the equating-test-total-test correlation was unavailable, the one available correlation for that equating was assigned a weight of 2 rather than 1. The grand total was thus 24 points.

These weights could not be applied to the "raw" equating indices because the indices are on different numerical scales. For purposes of forming the composite index, each of the indices was expressed on a three-point scale (1=low, 3=high). This scale was determined normatively by dividing the observed range for a given individual index into three equal intervals. The observed range was taken over all indices of the same type (i.e., standardized differences in means, variance ratios, correlations) for the SAT-V and the SAT-M separately. For example, the range for the standardized difference between equating means for the SAT-V was based on the minimum and maximum values observed for November and December at both equatings--a total of 60 observations.

Transformations were made to three of the individual equating indices for determining the boundaries of the three-point scale: (1) The absolute values rather than the signed values were used for the standardized differences between equating test means; (2) the ratios of the equating test variances were transformed to the natural log scale and then absolute values were determined, thus providing equivalent measures for values less than or greater than one; (3) the correlations were changed to normal deviates by the transformation

$$z = .5\log_e \frac{1 + r}{1 - r},$$

The ranges of the transformed variables were divided by three to obtain the cut points. The boundaries for the three-point scale represented equal intervals on the transformed scale but not necessarily on the original scale. The overall composite equating index was computed by taking the average of the weighted composite.

Trends in Equating Indices

The information on equating indices is provided in Figures 13 and 14 and in Tables B-14 to B-17 in Appendix B, which include the boundaries for the three-point scale expressed in the original metrics. To allow easier interpretation of trends in the data, the figures depict the absolute values of the standardized differences and the inverses of variance ratios less than 1 (e.g., .8 was plotted as 1.25).

Standardized differences between sample means. Figures 13 and 14 (Panels a and b) show the differences in standardized equating-test means between new-form and old-form samples. The plots for the SAT-V and the SAT-M are very similar--as one might expect when variables are correlated and when the samples from the same populations are used in equating. (The SAT-V and -M equating samples for a given test administration were essentially random samples from the population of test takers at that administration.) Therefore, whatever is said here applies to SAT-V and SAT-M equatings except where noted.

Of the November SAT-V or -M standardized differences, only those for 1982 exceeded the .25 cutoff used to decide between Tucker and Levine equating methods. Most of the values were less than .15, indicating that the differences between new- and old-form equating samples were relatively small in November. From 1979 to 1984 four of the SAT-V or -M standardized differences for the second equating tended to be larger than the other differences. Otherwise there is no indication of any differences on this index among the various equatings.

The standardized differences for December SAT-V or -M equatings, on the other hand, were much more variable, and eight SAT-V and four SAT-M values exceeded the .25 cutoff (see Tables B-14 and B-15). From 1980 on there were fairly large discrepancies in standardized means between the two sets of new-form-old-form equating samples. However, these discrepancies were due to differences between the equating samples for the first equatings that were smaller than the general trend, not to differences that were larger. There is no indication that later December equatings had larger standardized mean differences than earlier equatings.

Ratios of equating-test variances. The ratio of the variances of the new- and old-form samples on the equating test is another indicator used in equating. (See Panels c and d of Figures 13 and 14.) A ratio that is greater than or equal to 1.25 or is less than or equal to .80 suggests that Levine equating rather than Tucker equating should be used. For plotting purposes, values less than 1.00 were converted to values greater than 1.00 simply by dividing the original value into 1.00. This transformation permitted differences in the plots to be compared without having to consider scale direction. The trends for the SAT-V and the SAT-M tended to be somewhat different and thus are discussed separately.

The November and December SAT-V ratios of equating-test variances for new- and old-form samples all were less than the 1.25 cutoff used to decide between Tucker and Levine equating. The November and December ratios tended to fall between 1.00 and 1.10. The November 1982 and 1983 ratios were the two highest observed in November. Relatively large December ratios occurred in 1971, 1980, and 1982. The only evidence of any trend frcm 1970 to 1984 was the increase in the ratios for the November first equating from 1980 to 1983.

The ratios for the SAT-M fluctuated more than those for the SAT-V. As a result, larger differences occurred between the two new-form-old-form-sample ratios at any given administration. Differences were particularly small, however, in November from 1976 to 1979. None of the ratios came close to 1.25, but a number were close to 1.15. The larger ratios tended to occur throughout the 1970-84 period in both November and December. No systematic trends were evident from the data.

Correlations between equating tests and total tests. The plots for the correlations between equating and total tests (see Panels e and f of Figures 13 and 14) are more difficult to interpret because two correlations exist for each equating: one for the new-form sample and one for the old-form sample. Because three of the early 1970 equatings used internal equating tests for new-form samples, which caused the correlations to be spuriously high, the affected correlations were adjusted to make them comparable to the correlations for external equating tests. The old-form sample for the November 1974 SAT-V equating involved an equating test that was partly internal and partly external because only some of its items were used in the operational test. This value could not be corrected and stands out in the plot because it is spuriously high. This value should be ignored in assessing trends. All other correlations are comparable.

The correlations between SAT-V equating and total tests were slightly higher for November than for December. They tended to vary between .86 and

.88 for November and between .85 and .87 for December. No effect was noticeable due to the introduction of the shortened SAT in 1974, which if anything should have caused the correlations to be slightly lower for SAT-V new-form samples. The plot shows that the correlations for both old- and new-form samples were up slightly in November 1974 and down slightly in December 1974. One notes little difference between the correlations from 1974 on and the correlations before 1974 for either November or December. In November one of the old-form samples tended to have higher correlations than other three samples in 1980, 1981, 1982, and 1984; the other correlations seemed slightly depressed. In December one of the old-form samples tended to have higher correlations than the other samples in 1970-73.

In the case of the SAT-M, the correlations were somewhat lower than those for the SAT-V and more variable. The fluctuation occurred throughout the 1970-84 period. The average correlation was approximately .85 in both November and December. Especially large differences among the four equating-test-total-test correlations from given administrations occurred in November 1978 and 1981. Such differences can affect the quality of equating. No systematic trends were evident in the data.

Differences between equating lines. Absolute values of the differences between scaled scores at the raw-score midpoints for SAT-V and SAT-M were also plotted (see Panel g of Figures 13 and of Figure 14). In a more comprehensive analysis, differences across the entire scale would be considered;

such an analysis was beyond the scope of this study. Differences between equating lines, particularly differences in the densest part of the observed-score distributions are important in evaluating the consistency of equating. Ideally, the equating lines would agree. The differences at the midpoints provide evidence, albeit limited, of the consistency of the score conversions resulting from the two individual equatings.

The differences between equating lines for SAT-V equatings were quite variable, ranging from near 0 to over 15. The differences for November were relatively large in 1974 and 1975; those for December were relatively large in 1973 and 1980. The November equatings manifested a run of small differences in 1970-73 and in 1978-82. Otherwise no trends were evident.

The differences between equating lines for the SAT-M were much less variable than those for the SAT-V. The only differences that stood out in the plot of November and December differences were those for the December 1973 and 1974 forms. Interestingly, the November equatings tended to have slightly larger differences between equating lines than did the December equatings. Only five of the differences for December exceeded the November differences. No particular trends were noticeable in the data.

Equating Composite. The equating composite provides an overall index that combines information from the individual indices. While the particular composite used here was arbitrary, it was based on some logical considerations as to which of the individual indices mattered most in an equating. This composite index reflects only some of the variables that could be used to compute such an index and thus should not be interpreted as a measure of equating quality. Nevertheless, as an overall equating index, the equating composite provides a useful way of taking all of the variables into account at the same time. Panel h of Figures 13 and Panel h of Figure 14 show the plots of the SAT-V and SAT-M equating composites.

The composite equating indices for SAT-V equatings varied considerably and were relatively low in 1974-75 and in 1982-84 for November equatings, and in 1973, 1980, and 1981 for December equatings. The relatively high period, on the other hand, occurred in 1970-73 and 1977-81 for November equatings and 1970-73 and 1979 for December equatings. Only in 1983 was a composite index from December higher than that from November. The SAT-V December 1973 equating had the lowest composite index (1.6) and the SAT-V November 1978 equating, the highest (3.0). Although the composite indices for November were relatively low in 1982-84, the equating composites for December SAT-V equatings in 1982-84 were as high as many of the other December composites. Therefore, the data did not provide evidence of a decrease in the equating composite after the change in SAT-V item-difficulty specifications and the introduction of IRT equating in January 1982. No systematic patterns were evident in the composite equating indices for SAT-V equatings from 1970 to 1984.

The composite indices for SAT-M equatings were less variable than those for SAT-V equatings. The composite indices for December SAT-M equatings were similar to those for November except for the 1973-75 period, when the indices were low. The highest composite index (2.9) occurred in November 1977 and in December 1970, and the lowest (1.8) in December 1973. Although the equating composites for the more recent SAT-M equatings were slightly lower than for earlier equatings, the trends in the composite indices appeared to be unsystematic and unpredictable.

Of interest is the comparison of the composite equating indices for the different periods: 1970-73, 1974-77, 1978-81, and 1982-85. The means for these periods are shown under the column headed "Period Avg." in Tables B-16 and B-17. In the case of the SAT-V equatings, the periods with the highest average equating composite were the 1970-73 and 1978-81 periods for November. For the SAT-M equatings, the highest average composite came in the 1970-73 and 1974-77 periods, also for November. The overall equating composites for both the SAT-V and the SAT-M tended to be lower in December than in November, particularly in the 1970-73 and 1974-77 periods.

Although there were some equatings with relatively low composite equating indices, the evidence did not indicate a general decrease in the composite equating index from 1970 to 1984. Nor did the data indicate a decrease in the composite index due to the shortening of the SAT in 1974 or to the change in SAT-V statistical specifications in 1982 and the introduction of

IRT equating for both the SAT-V and the SAT-M.

Changes in Equating Methods and Trends in Predictive Validity

The primary consideration in this section is to assess how changes in equating might have affected correlations between the SAT and a criterion. In general, equating, by making more comparable scores from different administrations, would be expected to improve validity relative to the validity of unequated scores. That is, across test administrations, one would expect correlations based on scaled scores to be higher than correlations based on raw scores.

There is a condition under which equating does not affect predictor-criterion correlations: when all of the scores come from a single test administration, and linear equating is used. Then correlations based on scaled scores are no different from correlations based on raw scores, for correlations are not affected by linear transformations. There is also a condition under which equating attenuates validity: when scaled scores are curvilinear transformations of raw scores but the criterion is linearly related to raw scores. In general, however, curvilinear equating would be expected to "straightens out" a curvilinear relation with the criterion that is introduced at the raw score level, as when a new form is much easier or much harder than its predecessors.

As Tables 2 and 3 show, more of the SAT scores for a given college-bound senior cohort came from the November administration than from any other administration, and over half came from the November and December administrations. Given this condition, one would not expect equating to influence validity very much, even if equating produced scaled scores that were slightly off scale at different administrations, particularly for forms with conversions established by linear methods. The equatings for forms administered at times other than November and December would presumably influence validity only marginally because only a small proportion of individuals in the validity-study samples would have taken any one of these forms.

Of course, to account for increases or decreases in validity, equating procedures would have to have affected scores differently over time. Presumably, changes in equating procedures would have occurred during times of increasing validity and also during times of decreasing validity. The evidence presented in the section on equating changes does not indict equating methods as a cause of validity decline. The equating indices, including the equating composite, showed little difference from the 1970-74 period to the later periods. Few consistent patterns were discernible in the data on equating indices. Still, it may be instructive to review other evidence that bears on the relationship between equating and validity.

Comparisons of Results from Different Equating Methods

When the shortened SAT was introduced in the fall of 1974, linear rather than curvilinear equating was used. At that time the number of SAT-V items was reduced from 90 to 85, while the number of SAT-M items remained at 60. A curvilinear equating procedure might have been appropriate for the SAT-V then--at least during the time test forms from the 1970-73 period provided the old-form equating data. Curvilinear procedures were not introduced until January 1982, however, when the verbal item-difficulty specifications were changed. (The computing power necessary to use IRT equating with the SAT was not available in the late 1970's; otherwise IRT equating might have been used earlier.) Equipercentile equating results, which express curvilinear relationships between new- and old-form scores, were routinely available during the 1970-85 period and may be compared with operational equating results.

Did the decisions to continue to use linear equating in 1974 and to switch to IRT equating in 1982 affect reported scores in ways that could have influenced predictive validity? Comparing the results of different equating methods helps answer this question. Two comparisons were made in this study regarding the effect of choice of equating methods in 1974 and 1982 on reported scores. One was based on an index that measured the discrepancy between the operational equating line and the equipercentile equating line at the midpoint of the raw score range. The other was based on indices that measured the discrepancy of equating results across the entire score range for equatings in November 1974, December 1974, and January 1982.

Comparison of operational and equipercentile equating results at raw-score midpoints. During the 1970-74 period, before the SAT was shortened, linear methods were used to equate SAT scores. The use of linear methods at that time was due primarily to the fact that test developers assembled the test forms administered during the period to the same statistical specifications, thus ensuring essentially parallel forms. It was due also to the lack of a fully satisfactory curvilinear equating procedure and the unavailability of a computer system that accommodated curvilinear equating. Although equipercentile equating through an anchor test was performed in addition to linear equating, it was used only as a check on the curvilinearity of equating and as a basis for empirical "doglegs," which were linear line segments covering a small part, usually the upper end, of the score range.

Despite changes in statistical specifications for SAT-V and SAT-M in 1974 when the shortened SAT was introduced, linear methods continued to be used. The specified distribution of item difficulties for SAT-V became more like that of the period before 1966 except for an increase in the standard deviation of item difficulties. The bimodal distribution included more items at delta levels greater than or equal to 15 to ensure good measurement at the upper end of the scale despite the decrease in test difficulty. The distribution of SAT-M item difficulties shifted downward slightly to make the test less difficult but otherwise looked very much like the distribution for the earlier period.

Despite these changes in specifications, linear methods continued to be used to equate SAT scores until January 1982, when IRT methods were introduced. The change in SAT-V statistical specifications in 1982, which called for fewer difficult items but the same mean item difficulty as before, was expected to produce curvilinear relationships between scores on new forms and scores on previous forms. IRT equating is an equating method that not only permits curvilinear relationships but also, as a true score method, can adjust for relatively large ability differences in equating samples.

It is possible that the use of linear rather than curvilinear equating for the shortened SAT could have affected reported scores, especially when new forms were equated to forms administered prior to the fall of 1974. It is likewise possible that the switch to IRT equating in 1982 could have affected reported scores in ways that influenced predictive validity.

Linear and curvilinear equating results exist for all equatings conducted from 1970 to 1985. During these years Tucker equating, Levine equating, and equipercentile equating through an anchor test were performed routinely. Equipercentile equating produces curvilinear results, whereas the other two methods produce linear results. Theoretically, equipercentile equating and other curvilinear equating methods will result in a line that is essentially linear if linear equating is appropriate.

Equipercentile equating through an anchor consists of two separate equipercentile equatings: one linking scores on the new form to scores on the equating test and one linking the equating test scores to scores on the old form. This type of equipercentile equating, which uses only the four separate marginal distributions on the anchor test and total tests, is theoretically inferior to equipercentile equating that uses information from the bivariate distributions of total-test and equating-test scores. Still, it is instructive to compare the operational equating results with the results that would have obtained had equipercentile equating been used. Here comparisons made use of data at the midpoint of the raw score range, where much of the data were relatively dense.

If curvilinear equating were appropriate in 1974, one would expect to see the following differences between the operational (linear) and equipercentile equating lines at the midpoint of the raw score range for the November and December SAT test forms administered from 1970 to 1981:

- Relatively large deviations for the equatings of shortened test forms that went back to old forms administered before the fall of 1974;
- Somewhat smaller deviations for the equatings of shortened test forms that went back to one old form administered before the fall of 1974 and one old form administered after the shortening of the test;
- Relatively small deviations for the equatings that went back to old test forms that were similar to the new form, as during the 1970-73 and 1978-81 periods.

Because the 60-minute SAT-M test forms, although somewhat easier than previous forms, contained 60 items like the previous forms, slightly smaller deviations than for SAT-V forms were expected to occur for equatings going back to forms administered prior to the fall of 1974.

If linear equating were appropriate in 1974, then one would expect to see the following:

- o relatively small deviations between the operational and equipercentile equating lines in the 1970-81 period;
- o no larger deviations for shortened forms that went back to one or two forms administered before 1974 than for forms that went back to old forms similar to the new form.

One would also expect to see relatively small deviations between the operational (IRT, sometimes in combination with other methods) and equipercentile equating results in the 1982-84 period regardless of whether linear or curvilinear equating were used from 1982 to 1984. Thus, these data did not address the appropriateness of IRT equating. The appropriateness of IRT equating is dealt with in the discussion of the data from the selected equatings in January 1982.

Table 36 gives the differences at the midpoint of the raw-score range for the November and December forms from 1970 to 1984. Table 37 summarizes the data in Table 36 by distinguishing between the equatings of new forms that went back to forms from an earlier period and those of new forms that went back to forms within the same period. These comparisons are referred to as "between" and "within" in the table. Equatings for which one old form came from the same period and one came from an earlier period are referred to as "mixed" comparisons. Because equating lines can deviate from one another in either direction, signs were ignored to compute the values in the summary table.

Two types of evaluations were appropriate here. One took into consideration the size of the deviations. The other took into consideration the deviations between periods compared with deviations within periods. Table 36 shows that, in general, departures from the equipercentile lines at the midpoints were small--most entries are less than 5.0. The maximum value is 7.3--for the December 1976 SAT-M form. The small sizes of the deviations suggest that at least in the 1970-81 period, linear equating was appropriate in most cases.

The comparisons of deviations for the different periods found no larger differences for deviations from between and mixed comparisons than for deviations from within comparisons. Interestingly, the departure from equipercentile equating results were about as large for IRT equating as for linear methods. Thus, between-method variation was as large when similar curvilinear methods were compared (1982-84) as when linear and curvilinear methods were compared. Contrary to expectations, the mean deviations of the equatings of the 1974 and 1975 November and December SAT-V forms were not as large as those for some of the within-period equatings. In the case of the SAT-M, the mean deviation for the November 1974 and 1975 forms exceeded any other mean deviations. On the other hand, in December the mean deviation for the 1974 between-period SAT-M equating was not as large as the within-period equatings in years 1978 to 1981. For both the SAT-V and the SAT-M, only the December SAT-V administrations showed the expected pattern of increasing mean deviations from within- to mixed- to between-period comparisons. The mean deviation for the 1975-77 or 1976-77 period fell between the within-period and between-period equatings in only two of the four cases.

The inconsistency of the data indicates that the decision to retain linear equating in 1974 probably had little effect on reported scores. It suggests, moreover, that even if linear methods were inappropriate in a few instances, the effects were small and probably inconsequential.

Comparison of linear and curvilinear equating results for selected forms. Other data for comparing operational equating results with those of other methods came from two 1974 administrations and the January 1982 administration. These "case studies" provided additional evidence regarding the appropriateness of linear equating methods in 1974 and provide fresh evidence about the appropriateness of IRT equating in 1982. IRT procedures, sometimes in combination with other methods, were used to equate SAT scores during the 1982-84 period. IRT procedures were introduced in January 1982 because of the change in SAT-V specifications and the expectation that the raw-score-to-scaled score relationships would be curvilinear. For the 1974 and 1982 equatings, operational equating lines--either linear or IRT conversions--were compared with "experimental" equipercentile or linear equating lines. These comparisons were based on information on scores throughout the range rather than on information at only one score point and thus provide more information than the analysis discussed in the previous section.

Tables 38 and 39 show the comparisons of the four SAT-V and the four

SAT-M forms. The tables report scaled scores for selected raw scores and provide summary information about differences between the lines. The index that tells most about overall differences in the lines is the root mean squared difference. It takes account of the average difference as well as differences throughout the score scale and is computed by the formula

$$\sqrt{(\text{Mean}_{\text{diff}}^2 + \text{SD}_{\text{diff}}^2)}.$$

This index scaled to a standard deviation of 110 for operational scores is probably the most informative of the various indices. It is more comparable than the unscaled root mean squared difference from equating to equating because it adjusts for differential variation from group to group. The tables give the correlations between operational and experimental scores as well as root mean squared differences.

Again two types of evaluation are meaningful here: one taking into consideration the size of the difference between linear and curvilinear equating lines and one taking into consideration the differences between these lines for forms that went back to similar old forms and forms that went back to dissimilar old forms. The latter comparison was possible only for the 1982 equatings.

The amount of curvilinearity in the equating results is a matter of interest, as it could affect the relationship of SAT scores with other measures, including freshman grade point average. If changes introduced into the SAT-V could have been accounted for by either IRT or linear equating, then the particular equating method used would not have affected test scores to any large extent.

The differences between linear and curvilinear equating results bear on the issue of whether the choice of equating method affected test scores. The experimental and operational conversions and the root mean squared differences were compared for the 1974 equatings and the 1982 equating for the SAT-V and SAT-M. For the 1974 equatings the experimental line was derived through equipercentile equating; for the 1982 equatings it was derived through linear equating.

The equipercentile equating results for the 1974 SAT-V and SAT-M forms, all of which were built to changed specifications, were similar to the linear operational results. The score conversions and the means and standard deviations of the score distributions were very similar. Also, the root mean squared differences for these four forms were similar in magnitude to those for the January 1982 SAT-V form built to previous specifications and less than the root mean squared differences for the January 1982 SAT-M forms, which were also built to previous specifications. Moreover, the correlations between linear and equipercentile scores were essentially .998 or .999. These data suggest that continuing to use linear equating in 1974 had little effect on reported scores.

The comparison of the SAT-V equating results for the January 1982 administration in particular allows an assessment of the effect of changing the verbal statistical specifications and switching equating methods. One of the SAT-V forms was built to the new statistical specifications and one was built to previous specifications. Presumably, if curvilinear equating procedures were needed, the equating results for the form built to the new specifications would deviate more from linearity than those for the other form. The results in Table 38 show, however, that the operational IRT conversions were more similar to the linear conversions for the form built to the new specifications than for the form built to previous specifications. In addition, the root mean squared difference between the IRT and linear equating results was 5.19 for Form 2, which was assembled to revised specifications, compared with 6.73 for Form 1. In both cases the scaled score means and standard deviations associated with the two equating lines differed by no more than a point or two. This comparison demonstrates that the change to IRT equating in 1982 had little effect on scaled-score conversions.

For the January 1982 SAT-M forms, which were built to the same specifications as their predecessor forms, the linear and IRT means were less than three points apart. The standard deviation for the IRT equating of Form 1 in January, however, was nearly four points higher than the standard deviation for the linear equating. The root mean squared difference for this form was also larger than that for the other January SAT-M form. The scaled score conversions at selected raw scores indicate that the IRT equating line was higher at the higher end of the score scale and lower at the lower end. It appears that this particular form, despite being assembled to the same statistical specifications as the other form, yielded scores that were slightly curvilinear.

In January 1982 the test scores produced by linear equating were similar to those produced by IRT equating for both SAT-V forms and for one of the SAT-M forms. Thus, the choice of equating method was of little consequence.

Effect of Curvilinear Equating on Predictive Validity

Little evidence exists of the direct effect of curvilinear equating on predictive validity. Data that were available on a sample of test takers who were included in the 1985 VSS validity-study samples, however, did permit a direct comparison of the validity of linear equating results and the validity of curvilinear equating results. A match of November 1984 testing program files with VSS files for the entering class of 1985 resulted in a sample of 59,383 test takers who had SAT scores, both raw (formula scores) and scaled, and freshman grade point averages.

The SAT-V equating for November 1984 was based on two IRT equating lines, and the SAT-M equating was based on one IRT line and three linear lines (see Table 33). Because IRT equating results in curvilinear conversions, both SAT-V and SAT-M scaled scores for the November 1984 administration had curvilinear relationships with raw scores. Given that correlations of raw scores are not affected by linear transformations, the correlations of raw scores with freshman grade point average are those that would have resulted from linear equating lines. Thus, one could compare the correlations of raw and scaled scores with freshman grade point average to assess the effects of using curvilinear conversions.

The raw- and scaled-score correlations with freshman grade point average were very similar. SAT-V raw scores correlated .3662 with freshman grade point average, while SAT scaled scores correlated .3650. The corresponding correlations for the SAT-M were .3445 and .3442. Thus, for this sample the use of linear conversions would have had little effect on predictive validity. The correlations are affected to an unknown extent by differences in the meaning of grades from one college to another. Such differences would attenuate the correlations of SAT-V and SAT-M with college grades across colleges, but would not be expected to affect the essential singularity of raw scores and scaled scores and the relationships of these scores with grade point averages.

The analysis of data from test takers from other test forms could yield different results. The findings reported here, however, coupled with the small differences found in most cases between linear and curvilinear equating results, suggest that the choice of choice of equating method had little effect on predictive validity.

Scale Stability

Perhaps the strongest evidence of the integrity of the equating process comes from the scale-stability studies that have been conducted. The latest study (McHale & Ninneman, 1990) covered the period 1973 to 1984 and thus is directly relevant to the time period addressed in the current study. Earlier scale stability studies (Stewart, 1966; Modu & Stern, 1975, 1977) focused on the periods 1948-63 and 1963-1973, respectively. The findings from the earlier studies were as follows:

1948-53: SAT-V scale 20-35 points higher in 1953;
SAT-M scale not investigated

1953-63: SAT-V scale stable;
SAT-M scale not investigated

1963-73; 1966-73: SAT-V scale 8-14 points higher in 1973;
SAT-M scale 17 points higher in 1973

The 1966 and 1975 studies utilized the anchor-test design, in which two equating sections previously administered with the old forms were administered with a new form of the SAT. The 1977 study collected data from a special SAT administration--the old and new forms were spiralled (form A, form B, form A, form B, etc.) in packets and administered in high schools. The 1975 and 1977 studies identified a likely bias in the anchor test design--equating results are slightly too high when new-form samples are of lesser ability than old-form samples, and slightly too low when they are of higher ability.

McHale and Ninneman took account of this possible bias by using not only the traditional anchor-test design but also a spiralled-section design. The anchor-test design involved the administration of verbal and mathematical equating sections given with three 1973 and one 1974 SAT. These sections were re-administered with two 1983 and two 1984 SAT forms. The spiralled-section design involved the administration of each of the SAT-V and the SAT-M operational sections of a 1974 SAT form in the variable section of a 1984 form.

In addition, the operational sections of a 1975 form were administered with a different 1984 form. The latter design is theoretically superior to the anchor test design in that its results come from essentially equivalent samples. Because a given sample took only one of the two sections of an old form, however, special statistical (section pre-equating and IRT) methods had to be applied to estimate results for a complete old form. Thus, both designs suffered to some extent from certain weaknesses. Of course, both designs assumed that the old testing material was appropriate for the new groups.

McHale and Ninneman concluded that the SAT-V scale was relatively stable from 1973 to 1984. They found, however, inconsistent results for SAT-M. The four anchor-test equatings suggested an upward drift of 6 to 13 points, whereas two spiralled equatings indicated a downward drift of 6 to 14 points. While this study showed a possible drift in the SAT-M scale between 1973 and 1984, the drift is at worst no more than one and a half scaled-score points a year. Scale shifts this small are unlikely to have any effect on validity.

REFERENCES

Angoff, W. H. (1956). A note on the estimation of nonspurious correlations. *Psychometrika, 21*, 295-297.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. R. Rubin (Eds.), *Test equating*. New York: Academic Press.

Angoff, W. H., Pomplun, M., McHale, F., & Morgan, R. (in press). Comparative study of factors related to the predictive validities of 1974-75 and 1984-85 forms of the SAT. In Willingham, W. W., Lewis, C., Morgan, R., & L. Ramist, *Predicting college grades: An analysis of institutional trends over two decades*. Princeton, NJ: Educational Testing Service.

Bloom, B. J. (Ed.). (1956). Taxonomy of educational objectives: the classification of educational goals. New York: David McKay.

Braswell, J., & Marco, G. L. (1974, March). *Proposed statistical specifications for SAT-V and SAT-M*. Unpublished memorandum, Educational Testing Service, Princeton, NJ.

Burton, N. W., Morgan, R., Lewis, C., & Robertson, N. J. (1989, March). *The predictive validity of SAT and TSWE item types for ethnic and gender groups*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

College Entrance Examination Board. (1984). *Taking the SAT*. New York: Author.

College Entrance Examination Board. (1988). *10 SATs* (3rd ed.). New York: Author.

Cruise, P. I., & Kimmel, E. W. (1990). *Changes in the SAT-Verbal: A study of trends in content and gender references, 1961-1987* (College Board Report No. 90-1 and ETS Research Report 89-17). New York: College Entrance Examination Board.

Donlon, T. F. (1984). The Scholastic Aptitude Test. In T. F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 37-67). New York: College Entrance Examination Board.

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.

Donlon, T. F., & Livingston, S. A. (1984). Psychometric methods used in the Admissions Testing Program. In T. F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 13-36). New York: College Entrance Examination Board.

Educational Testing Service. (1980). *Test sensitivity review guidelines*. Princeton, NJ: Author.

Educational Testing Service. (1989). *A reader's guide to Scholastic Aptitude Test (SAT) test analysis report* (Statistical Report No. 89-61). Princeton, NJ: Author.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Marco, G. L. (1973, November). *Preliminary data on a shortened SAT.* Unpublished memorandum, Educational Testing Service, Princeton, NJ.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). *A large-scale evaluation of linear and curvilinear equating models, Vol. I* (Research Memorandum No. 83-2). Princeton, NJ: Educational Testing Service.

McHale, F. J., & Ninneman, A. M. (1990). *The stability of the score scale for the Scholastic Aptitude Test from 1973 to 1984* (ETS Research Report 90-6). Princeton, NJ: Educational Testing Service.

McPeek, M. (1970, April). *Experimental pretests in RSA45 (November 1969).* Unpublished memorandum, Educational Testing Service, Princeton, NJ.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT score scale* (College Board Research and Development Report 74-75, No. 3). New York: College Entrance Examination Board.

Modu, C. C., & Stern, J. (1977). *The stability of the SAT-Verbal score scale.* In College Entrance Examination Board, *Appendixes to on further examination,* New York: Author.

Morgan, R. (1989). *Analyses of the trends in predictive validity of the SAT and high school grades from 1976 to 1985.* (College Board Report No. 89-7 and ETS Research Report No. 89-37). New York: College Entrance Examination Board.

Petersen, N. S. (1981, April). *Interim delta specifications for SAT-Verbal.* Unpublished memorandum, Educational Testing Service, Princeton, NJ.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: a comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Ramist, L. (1984). Predictive validity of the ATP tests. In T. F. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 141-170). New York: College Entrance Examination Board.

Ramist, L., & Weiss, G. E. (in press). The predictive validity of the SAT, 1964 to 1988. In Willingham, W. W., Lewis, C., Morgan, R., & L. Ramist, *Predicting college grades: An analysis of institutional trends over two decades.* Princeton, NJ: Educational Testing Service.

Schrader, W. B. (1973). *Validity of the quantitative comparison test* (Statistical Report No. 73-60). Princeton, NJ: Educational Testing Service.

Schrader, W. B. (1984). *Three studies of SAT-Verbal item types* (College Board Report No. 84-7 and ETS Research Report No. 84-33). New York: College Entrance Examination Board.

Stewart, E. E. (1966). *The stability of the SAT-Verbal score scale* (College Board Research and Development Report 66-7, No. 3). New York: College Entrance Examination Board.

Swineford, F. (1974). *The test analysis manual* (Statistical Report No. 74-06). Princeton, NJ: Educational Testing Service.

Walker, R. C. (1981). *A reader's guide to test analysis reports.* Princeton, NJ: Educational Testing Service.

Wright, D., Wright, N., & Weber, C. (1985). *Test analysis: College Board Scholastic Aptitude Test, December 1984 administration, 3GSA09* (Statistical Report No. 85-185). Princeton, NJ: Educational Testing Service.

Table 1. Average Adjusted Correlations of the SAT with College Freshman Grades for Classes Entering College from 1970 to 1985[a,b]

Year	SAT-V[c]	SAT-M[c]	Multiple SAT[c]
1970	.48	.46	.52
1971	.50	.51	.56
1972	.48	.47	.52
1973	.50	.50	.55
1974	.51	.53	.57
1975	.49	.51	.56
1976	.51	.51	.56
1977	.50	.51	.56
1978	.49	.50	.55
1979	.48	.49	.53
1980	.48	.50	.54
1981	.48	.49	.54
1982	.49	.50	.54
1983	.48	.49	.53
1984	.46	.48	.52
1985	.47	.47	.52

[a] The estimates were based on data from the College Board Validity Study Service (VSS). The correlations were estimated from data on 472 colleges that participated in the VSS more than once from 1970 to 1987 and reported college freshman grades on a scale of 0 to 4.

[b] Correlations were adjusted for restriction of range on SAT scores and high school record due to selectivity in college admissions and enrollments.

[c] The average correlations for the SAT-V and the SAT-M are the means of zero-order correlation coefficients with college freshman grades. The average correlations for the multiple SAT are multiple-correlation coefficients.

Table 2. Numbers of Test Takers, Scaled-Score Means, and Scaled-Score Standard Deviations for Various Groups of SAT-Verbal Test Takers by Year of High-School Graduation

	College-Bound Seniors[a]			November Administration Seniors[b]			December Administration Seniors[b]			Validity Study Service Entering Freshmen[c,d]		
Year	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
1971	1,116,311	455	111	369,119	462	107	288,296	459	110	32	471	84
1972	1,022,820	453	111	365,456	459	109	242,784	451	109	100	480	87
1973	1,014,853	445	108	363,056	452	107	214,583	438	106	88	460	85
1974	985,239	444	110	362,023	458	108	212,213	4-4	110	111	465	88
1975	996,428	434	109	379,580	447	109	209,400	424	107	130	463	88
1976	999,809	431	110	396,457	442	106	205,082	408	109	139	463	90
1977	979,396	429	110	404,099	442	106	185,096	406	105	151	449	89
1978	989,185	429	110	373,258	435	108	184,227	408	105	173	449	86
1979	991,617	427	110	332,338	436	106	173,669	400	104	185	448	87
1980	991,245	424	110	329,601	434	106	183,235	396	102	181	443	86
1981	994,046	424	110	288,513	432	107	200,864	403	102	162	446	84
1982	988,270	426	110	315,736	432	108	171,933	402	99	173	444	86
1983	962,877	425	109	382,735	434	104	158,574	396	101	146	448	85
1984	964,684	426	110	367,767	437	106	160,249	395	102	154	449	86
1985	977,361	431	111	381,474	440	107	162,126	403	101	143	459	85

[a] Except for 1971, data are from College Board national reports on college-bound seniors; the 1971 data are estimates from testing year data for 1969-70 juniors and 1970-71 seniors.

[b] The November and December senior data are from the fall of the year preceding the year of graduation.

[c] Based on data reported in Ramist (1984), pp. 162-163, expanded to include data for classes entering in 1983, 1984, and 1985

[d] The N's are the numbers of colleges utilizing the Validity Study Service; the means and standard deviations are the unweighted averages of entering-class means and standard deviations. (Summary data on individuals are not available.)

BEST COPY AVAILABLE

Table 3. Numbers of Test Takers, Scaled-Score Means, and Scaled-Score Standard Deviations for Various Groups of SAT-Mathematical Test Takers by Year of High-School Graduation

	College-Bound Seniors[a]			November Administration Seniors[b]			December Administration Seniors[b]			Validity Study Service Entering Freshmen[c,d]		
Year	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
1971	1,116,311	488	113	369,119	499	110	288,296	492	114	32	498	83
1972	1,022,680	484	115	365,456	487	115	242,784	484	113	100	509	88
1973	1,014,704	481	113	363,056	490	113	214,593	482	113	88	497	87
1974	985,115	480	116	362,023	493	115	212,213	469	114	111	500	91
1975	996,391	472	115	379,560	482	112	209,400	460	112	130	501	92
1976	999,776	472	120	396,457	482	121	205,082	455	116	139	503	94
1977	979,344	470	119	404,099	480	118	185,096	455	114	151	485	93
1978	989,049	468	118	373,258	475	114	184,227	444	115	173	483	90
1979	991,405	467	117	332,338	472	114	173,669	445	110	185	484	90
1980	991,056	466	117	329,601	477	114	183,235	444	111	181	480	89
1981	993,672	466	117	288,513	474	114	200,864	445	112	162	486	87
1982	987,942	467	117	315,736	472	114	171,933	447	113	173	479	90
1983	962,542	468	119	362,735	473	114	158,574	443	109	146	487	89
1984	964,685	471	119	367,767	477	113	160,249	446	111	154	490	89
1985	977,361	475	119	381,474	485	115	162,126	452	113	143	498	87

[a] Except for 1971, data are from College Board national reports on college-bound seniors; the 1971 data are estimates from testing year data for 1969-70 juniors and 1970-71 seniors.

[b] The November and December senior data are from the fall of the year preceding the year of graduation.

[c] Based on data reported in Ramist (1984), pp. 162-163, expanded to include data for classes entering in 1983, 1984, and 1985

[d] The N's are the numbers of colleges utilizing the Validity Study Service; the means and standard deviations are the unweighted averages of entering-class means and standard deviations. (Summary data on individuals are not available.)

Table 4. Numbers of Items Specified Within Various Classifications for SAT-Verbal Tests from 1961 to the Present[a]

Item Type	Classification	Jan. 1961- Sept. 1974	Oct. 1974- Sept. 1978	Oct. 1978- Present
Sentence Completions	Content			
	Aesthetics/philosophy	4	4	4
	World of practical affairs	5	4	4
	Science	5	4	4
	Human relationships	4	3	3
	Structure			
	One missing word	6[b]	3-7	3-7
	Two missing words	12[b]	8-12	8-12
	(Total)	(18)	(15)	(15)
Antonyms	Content			
	Aesthetics/philosophy	4	6	6
	World of practice affairs	5	6	6
	Science	5	7	7
	Human relationships	4	6	6
	Generality of Required Distinction			
	General definition	10[b]	12-16	12-16
	Fine distinction	8[b]	9-13	9-13
	Structure			
	Single words	12[b]	13-17	15[b]
	Phrases	6[b]	8-12	10[b]
	Part of Speech Used			
	Verb	5[b]	4-10	4-10
	Noun	5[b]	4-10	4-10
	Adjective	8[b]	8-14	8-14
	(Total)	(18)	(25)	(25)
Analogies	Content			
	Aesthetics/philosophy	5	5	5
	World of practical affairs	5	5	5
	Science	5	5	5
	Human relationships	4	5	5
	Abstraction of Terms			
	Concrete	6[b]	4-8	4-8
	Abstract	5[b]	4-8	4-8
	Mixed	8[b]	6-10	6-10
	Independence of Stem and Key			
	Independent	13[b]	11-15	11-15
	Overlapping	6[b]	5-9	5-9
	(Total)	(19)	(20)	(20)
Reading Comprehension	Content			
	Narrative	5	5	2-5
	Biological science	5	0-5[d]	2-5
	Physical science	5	0-5[d]	2-5
	Argumentative	5	5	2-5
	Humanities	5	5	2-5
	Synthesis	5	0	0
	Social studies	5	5	2-5
	Functional Skill			
	Main idea	7	5	5
	Supporting idea	7	5	5
	Inference	12	9	9
	Application	3	2	2
	Evaluation of logic	3	2	2
	Style and tone	3	2	2
	(Total)	(35)	(25)	(25)

[a] The specifications applied to any new form administered during the indicated periods.

[b] No explicit deviations were listed as part of the specifications, but there were at times deviations as large as 4 in either direction.

[c] Beginning in December 1977, one of the reading passages was required to have a minority-group orientation.

[d] Only one science passage was permitted on the test.

Table 5. Actual Ranges in the Numbers of Items Within Various Classifications for November and December SAT-Verbal Test Forms from 1970 to 1984

Item Type	*Classification*	*1970-1973*	*1974-1977*	*1978-1984*
Sentence Completions	Content			
	Aesthetics/philosophy	4-5	3-5	4-4
	World of practical affairs	5-6	4-5	4-5
	Science	5-5	3-4	3-4
	Human relationships	3-4	3-3	3-3
	(Total)	(18)	(15)	(15)
Antonyms	Content			
	Aesthetics/philosophy	3-5	5-6	4-7
	World of practice affairs	4-7	6-7	5-7
	Science	4-5	6-7	4-8
	Human relationships	3-5	5-6	5-9
	(Total)	(18)	(25)	(25)
Analogies	Content			
	Aesthetics/philosophy	4-6	5-5	4-6
	World of practical affairs	4-6	5-6	5-7
	Science	5-6	4-5	4-6
	Human relationships	3-5	4-5	4-6
	(Total)	(19)	(20)	(20)
Reading Comprehension	Content			
	Narrative	5-5	5-5	3-5
	Biological science	5-10	5-5[a]	2-5
	Physical science	0-5	0-0[a]	3-5
	Argumentative	5-10	5-5	3-5
	Humanities	5-5	5-5	3-5
	Synthesis	0-5	0-0	0-0
	Social studies	5-5	5-5	3-5
	Functional Skill			
	Main idea	2-8	3-6	1-6
	Supporting idea	4-10	4-10	2-8
	Inference	11-16	6-10	8-11
	Application	1-4	1-3	1-3
	Evaluation of logic	1-6	1-3	1-4
	Style and tone	2-5	2-2	0-3
	(Total)	(35)	(25)	(25)

[a] Only one science passage was permitted on the test.

BEST COPY AVAILABLE

Table 6. Numbers of Items Specified Within Various Classifications for SAT-Mathematical Test Forms from 1969 to the Present[a]

Item Type	*Classification*	*Nov. 1969- Sept. 1974*	*Oct. 1974- Dec. 1975*	*Jan. 1976- Sept. 1981*	*Oct. 1981- Present*
Regular Mathematics	Arithmetic	13	12-13	12-13	12-13
	Algebra	11	11	11	11
	Geometry	13	11	11	11
	Miscellaneous	5	5-6	5-6	5-6
	(Total)	(42)	(40)	(40)	(40)
Data Sufficiency	Arithmetic	4-5			
	Algebra	4-5			
	Geometry	6-7			
	Miscellaneous	3-4			
	(Total)	(18)			
Quantitative Comparisons	Arithmetic		6	6	6
	Algebra		6	6	6
	Geometry		5-6	5-6	5-6
	Miscellaneous		2-3	2-3	2-3
	(Total)		(20)	(20)	(20)
All	Setting				
	Concrete	11-31	11-31	11-31	11-21
	Abstract	29-49	29-49	29-49	39-49
	Ability				
	Recall factual knowledge (Level 0)	0	0-3	2-21	0-1
	Perform math manipulations (Level 1)	0-3	4-13	0-3	0-3
	Solve routine problems (Level 2)	0-5	8-17	0-5	0-10
	Demonstrate comprehension of math ideas and concepts (Level 3)	22-30	22-31	22-41	22-43
	Solve nonroutine problems requiring insight or ingenuity (Level 4)	10-18	10-19	10-29	10-31
	Apply "higher" mental processes to mathematics (Level 5)[a]	20-28	7-16	7-26	7-28
	(Total)	(30)	(60)	(60)	(60)

[a] The specifications applied to any new form administered during the indicated periods.

Table 7. Actual Ranges in the Numbers of Items Within Various Classifications for November and December SAT-Mathematical Test Forms from 1970 to 1984

Item Type	*Classification*	*1970-1973*	*1974-1975*	*1976-1980*	*1981-1984*
Regular Mathematics	Arithmetic	13-13	12-12	12-13	12-13
	Algebra	11-13	11-11	10-11	11-11
	Geometry	12-13	11-12	11-11	11-11
	Miscellaneous	4-5	5-6	5-6	5-6
	(Total)	(42)	(40)	(40)	(40)
Data Sufficiency	Arithmetic	4-5			
	Algebra	3-5			
	Geometry	5-8			
	Miscellaneous	3-4			
	(Total)	(18)			
Quantitative Comparisons	Arithmetic		6-7	6-7	6-6
	Algebra		5-6	5-6	6-6
	Geometry		5-6	5-6	5-6
	Miscellaneous		2-3	1-3	2-3
	(Total)		(20)	(20)	(20)
All	Setting				
	Concrete	12-24	13-16	10-16	14-19
	Abstract	36-48	44-47	44-50	41-46
	Ability				
	Solve routine problems (Levels 0, 1, and 2)	7-10	14-20	10-18	11-17
	Demonstrate comprehension of math ideas and concepts (Level 3)	21-26	22-25	23-32	24-31
	Apply "higher" mental processes to math (Levels 4 and 5)	26-31	16-24	16-22	17-21
	(Total)	(60)	(60)	(60)	(60)

Table 8. Numbers of SAT-Verbal and SAT-Mathematical Items and Other Material Contained in Student Booklets De3cribing the SAT

		Sample Questions					
		Solution Method Explained		*Solution Method Not Explained*			
Testing Year(s)	*Total Number of Pages*	*Verbal*	*Math*	*Verbal*	*Math*	*Review of Basic Algebra and Geometry*	*Comments*
1970-71	55	16	17	57	36	No	Quantitetive Comperizons added to explained items (1970-71 to present)
1971-72	55	16	16	57	36	No	
1972-73 to 1973-74	15	10	7	0	0	No	Bulletins meiled to etudents with regietretion msterisls (1972-73 to 1973-74)
1974-75 tc 1975-76	12	4	8	21	8	No	
1976-77	15	8	7	15	14	No	
1977-78	16	0	0	30	30	No	
1978-79 to 1981-82	48	21	17	85	60	Yee	Expended bulletin introduced with full-length test (1978-79 to preeent)
1982-83 to 1984-85	62	23	18	85	60	Yes	

Table 9. Summary of Changes Made to SAT Item Types, Content, and Test Format from March 1970 to January 1985

Beginning Date	*Change*
March 1973	One minority-relevant reading passage included in at least one SAT-V form administered during the testing year
October 1974	Two 30-minute SAT-V sections (40 and 45 items, respectively) introduced in place of one 45-minute section (50 items) and one 30-minute section (40 items)
	Two 30-minute SAT-M sections (25 and 35 items, respectively) introduced in place of one 45-minute section (35 items) and one 30-minute section (25 items)
	30-minute Test of Standard Written English (50 items) introduced and administered in test booklet with the SAT
	Number of SAT-V items of a particular item type changed: o Number of Antonyms increased from 18 to 25 o Number of Analogies increased from 19 to 20 o Number of Sentence Completions reduced from 18 to 15 o Number of Reading Comprehension passages reduced from 7 to 5; number of Reading Comprehension items reduced from 35 to 25.
	Length and content of reading passages altered: o Total words in reading passages reduced from a maximum of 3,500 to 2,000-2,250; o Deletion of Synthesis and one of two science passages (Biological and Physical Science at the discretion of test assembler)

Table 9. (Continued)

Beginning Date	*Change*
	Reading (based on Reading Comprehension and Sentence Completion items) and Vocabulary (based on Antonym and Analogy items) subscores introduced: o Reading and Vocabulary items required to have similar mean item difficulties and standard deviations of item difficulties o Difficult vocabulary not used in Sentence Completion items o Number of more difficult Sentence Completion items increased.
	Number of SAT-M items of a particular item type changed: o Number of Regular Mathematics items reduced from 42 to 40 o 20 Quantitative Comparison items added o 18 Data Sufficiency items deleted
	Six (rather than one of two) fixed section orders used at each test administration
November 1975	To attempt to reduce speededness: o 10 Reading Comprehension items (based on two passages) moved from end to middle of SAT-V 40-item section, and 15 Reading Comprehension items (based on three passages) moved from the middle to the end of SAT-V 40-item section o 20 Quantitative Comparison items moved to middle of SAT-M 35-item section
1977-78	Slight reduction in the number of SAT-M items requiring a more complex knowledge of geometry
	Virtual elimination of the generic "he" from the SAT-V
December 1977	One minority-relevant reading passage included in each new form of the SAT-V

Table 9. (Continued)

Beginning Date	*Change*
October 1978	Number of Reading passages increased from five to six: o Three 200-250 word passages replaced two 400-450 word passages o Second science passage returned to test o Two to four rather than five items used for each shorter passage One of two fixed section orders used at each administration
1979-80	Seven rather than five new forms produced each year to fulfill the requirements of test disclosure
1980	Test sensitivity guidelines implemented; tests reviewed to eliminate any material offensive and patronizing to females and minority groups; representation in test items of contributions of females and minority groups to American society; improvement in the ratio of male-to-female references
October 1980	One of three fixed section orders used at each administration
1981-82	Nine or ten new forms produced each year to fulfill the requirements of test disclosure

Table 10. Correlations Among Item Types for SAT Test Forms Administered in March 1971 and in November and December from 1981 to 1984[a]

Item Type	No. of Items	SC	ANT	ANA	RC	RM	DS	QC
				March 1971 (N = 865)				
Sentence Completions (SC)	18	.75	.96	.96	.96	.73	.71	---
Antonyms (ANT)	18	.73	.76	.95	.89	.70	.67	---
Analogies (ANA)	19	.71	.71	.73	.94	.84	.78	---
Reading Comprehension (RC)	35	.76	.71	.73	.83	.74	.72	---
Regular Math (RM)	42	.60	.57	.68	.63	.89	.92	---
Data Sufficiency (DS)	18	.54	.50	.58	.57	.75	.75	.92[b,c]
Quantitative Comparisons (QC)	60	---	---	---	---	---	.76[b]	.91[b]
				November 1981-84 (N = 1490 - 1805)				
Sentence Completion (SC)	15	.66-.72	.96-.98	.91-.97	.93-.95	.66-.73	---	.66-.73
Antonyms (ANT)	25	.69-.73	.75-.79	.92-.95	.87-.89	.65-.68	---	.65-.69
Analogies (ANA)	20	.66-.71	.71-.73	.76-.78	.87-.89	.68-.72	---	.68-.74
Reading Comprehension (RC)	25	.68-.72	.68-.71	.67-.71	.81-.82	.60-.73	---	.60-.73
Regular Math (RM)	40	.50-.57	.53-.56	.55-.50	.59-.61	.87-.88	---	.98-.99
Quantitative Comparisons (QC)	20	.48-.54	.50-.53	.51-.58	.55-.58	.80-.82	---	.77-.79
				December 1981-84 (N = 1505 - 1910)				
Sentence Completion (SC)	15	.66-.72	.93-.96	.93-.96	.89-.93	.63-.71	---	.67-.71
Antonyms (ANT)	25	.68-.70	.74-.78	.92-.94	.87-.89	.61-.69	---	.63-.71
Analogies (ANA)	20	.63-.72	.68-.70	.71-.78	.86-.89	.70-.72	---	.71-.75
Reading Comprehension (RC)	25	.65-.70	.67-.68	.64-.89	.78-.80	.67-.72	---	.68-.71
Regular Math (RM)	40	.49-.56	.50-.55	.56-.60	.56-.59	.87-.89	---	.97-.99
Quantitative Comparisons (QC)	20	.47-.53	.48-.55	.53-.57	.52-.57	.78-.84	---	.73-.81

[a] The diagonal elements of the correlation matrices are internal-consistency reliability coefficients. Entries above the diagonals are correlations corrected for attenuation.

[b] The data for Quantitative Comparisons are based on 55 items administered in the variable section of the December 1970 SAT; the data for Data Sufficiency are based on 18 items administered in an operational section of the December 1970 SAT-M.

[c] Corrected using the Data Sufficiency reliability from the March 1971 administration.

Table 11. Test-Analysis Sample Sizes, Means, and Standard Deviations of Scaled Scores for November and December SAT Test Forms from 1970 to 1984

Year	Sample N[a]	SAT-Verbal		SAT-Mathematical	
		Mean	SD	Mean	SD
		November Administrations			
1970	2000	458	106	495	108
1971	2345	458	108	486	113
1972	1995	454	106	490	113
1973	1995	458	108	491	111
1974	1815	446	107	484	111
1975	1895	440	106	479	120
1976	1745	442	109	477	117
1977	1685	439	105	476	112
1978	1685	435	103	471	111
1979	1930	435	103	475	113
1980	1695	428	106	473	115
1981	1805	430	108	470	112
1982	1610	434	108	473	116
1983	1490	440	108	477	113
1984	1555	443	105	491	116
		December Administrations			
1970	2500	452	111	487	115
1971	2000	448	108	481	110
1972	1935	436	105	479	111
1973	1750	434	111	466	113
1974	1765	424	106	457	113
1975	1830	412	110	456	113
1976	1560	408	108	456	111
1977	1895	413	108	447	114
1978	1815	396	105	444	110
1979	2360	395	105	442	113
1980	1850	410	102	456	114
1981	1505	404	103	453	116
1982	1910	398	101	447	111
1983	1505	402	104	454	114
1984	1515	412	104	463	118

[a] From 1970 to 1980 the samples were statistically representative of the total population; from 1981 to 1984 samples were selected from junior and senior test takers.

BEST COPY AVAILABLE

Table 12. Statistical Specifications for SAT-Verbal and SAT-Mathematical Test Forms from 1966 to the Present[a]

Item Difficulty (Equated Delta)	SAT-Verbal			SAT-Mathematical	
	Aug. 1966-Sept. 1974[b]	Oct. 1974-Jan. 1982[c]	Jan. 1982-Present[c]	Aug. 1966-Sept. 1974	Oct. 1974-Present
≥ 18	0	0	0	3	3
17	2	2	0	4	4
16	4	4	2	4	4
15	8	10	6	4	4
14	10	10	14	5	4
13	10	6	10	5	4
12	10	6	8	5	4
11	10	6	7	8	8
10	10	8	7	8	8
9	8	8	10	7	8
8	7	10	8	4	5
7	6	8	6	2	1
6	3	4	4	1	2
≤ 5	2	3	3	0	1
Number of Items	90	85	85	60	60
Mean Delta	11.7	11.4	11.4	12.5	12.17-12.27
SD Delta	2.9	3.3	3.0	3.1	3.1-3.3
Mean Biserial r[d]	.42 (.47)	.43 (.48)	.41-.45 (.46-.50)	.47 (.53)	.47 (.53)

[a] The statistical specification applied to any new form administered during the indicated periods.

[b] From August 1966 to July 1967 the statistical specifications for SAT-V were as follows: Mean Delta - 11.8, SD Delta - 3.0, Mean Biserial r - .42.

[c] One of the two January 1982 forms was assembled to the specifications for the 1974-81 period.

[d] The mean biserial r is specified in terms of pretest items, which are not included in the total-score criterion. The equivalent means for a total-score criterion that includes the item, given in parentheses, are .05 higher for the SAT-V and .06 higher for the SAT-M.

Table 13. Specified and Actual Item Statistics for November and December SAT-Verbal Test Forms from 1970 to 1984

	Specified			November Actual			Decembe	
Year	Mean Equated Delta	SD Equated Delta	Mean Biserial r[a]	Mean Equated Delta	SD Equated Delta	Mean Biserial r	Mean Equated Delta	SD Equat Delt
1970	11.7	2.9	.47	11.9	2.8	.46	11.9	2.8
1971	11.7	2.9	.47	11.8	2.9	.46	11.8	2.9
1972	11.7	2.9	.47	11.5	2.9	.47	11.8	3.0
1973	11.7	2.9	.47	11.7	3.1	.47	11.7	3.1
1974	11.4	3.3	.48	11.5	3.3	.46	11.5	3.2
1975	11.4	3.3	.48	11.4	3.4	.48	11.4	3.2
1976	11.4	3.3	.48	11.3	3.4	.49	11.4	3.1
1977	11.4	3.3	.48	11.3	3.3	.46	11.3	3.1
1978	11.4	3.3	.48	11.4	3.1	.47	11.4	3.4
1979	11.4	3.3	.48	11.4	3.1	.46	11.5	3.3
1980	11.4	3.3	.48	11.1	3.3	.49	11.3	3.2
1981	11.4	3.3	.48	11.2	3.2	.50	11.3	3.4
1982	11.4	3.0	.46 - .50	11.5	3.0	.51	11.3	2.9
1983	11.4	3.0	.46 - .50	11.4	3.0	.50	11.2	2.8
1984	11.4	3.0	.46 - .50	11.3	3.1	.49	11.4	2.9

[a] Specified in terms of final-form items, which are included in the total-score criterion

Table 14. Specified and Actual Item Statistics for November and December SAT-Mathematical Test Forms from 1970 to 1984

	Specified			November Actual			December	
Year	Mean Equated Delta	SD Equated Delta	Mean Biserial r[a]	Mean Equated Delta	SD Equated Delta	Mean Biserial r	Mean Equated Delta	SD Equated Delta
1970	12.5	3.1	.53	12.3	3.0	.52	12.3	3.0
1971	12.5	3.1	.53	12.2	3.0	.54	12.3	3.0
1972	12.5	3.1	.53	12.4	3.1	.58	12.3	3.1
1973	12.5	3.1	.53	12.6	3.5	.54	12.5	3.0
1974	12.2 - 12.3	3.1 - 3.3	.53	12.5	3.6	.53	12.1	3.0
1975	12.2 - 12.3	3.1 - 3.3	.53	11.8	3.3	.58	11.8	3.1
1976	12.2 - 12.3	3.1 - 3.3	.53	12.1	3.2	.56	12.0	2.9
1977	12.2 - 12.3	3.1 - 3.3	.53	12.2	3.2	.54	12.4	3.5
1978	12.2 - 12.3	3.1 - 3.3	.53	12.1	3.1	.52	12.2	3.3
1979	12.2 - 12.3	3.1 - 3.3	.53	12.3	3.3	.54	12.2	3.2
1980	12.2 - 12.3	3.1 - 3.3	.53	12.2	3.1	.54	12.2	3.1
1981	12.2 - 12.3	3.1 - 3.3	.53	12.1	3.5	.56	12.4	2.9
1982	12.2 - 12.3	3.1 - 3.3	.53	12.3	3.3	.55	12.1	3.1
1983	12.2 - 12.3	3.1 - 3.3	.53	12.1	3.4	.55	12.0	3.2
1984	12.2 - 12.3	3.1 - 3.3	.53	12.6	3.3	.54	12.2	3.5

[a] Specified in terms of final-form items, which are included in the total-score criterion

Table 15. Mean Numbers of Items by Item Difficulty (Equated Delta) for November and December SAT-Verbal Test Forms Within Specified Periods from 1970 to 1984[a]

Equated Delta	*1970-1973*	*1974-1977*	*1978-1981*	*1982-1984*
		November Administrations		
≥ 18	0.0	0.0	0.3	0.3
17	1.8	1.5	0.5	0.0
16	5.3	5.5	3.0	2.0
15	5.8	9.0	11.3	7.0
14	10.3	9.0	8.1	13.0
13	9.0	7.8	9.1	9.7
12	10.5	4.5	6.3	7.4
11	11.8	7.3	5.5	11.1
10	10.8	7.8	7.8	5.7
9	7.0	8.0	8.8	8.4
8	7.3	7.8	8.3	7.4
7	6.3	8.5	9.6	5.7
6	2.8	5.3	3.0	3.0
≤ 5	1.8	3.3	3.5	4.3
Total[b]	90.5	85.3	85.1	85.0
		December Administrations		
≥ 18	0.3	0.0	0.0	0.0
17	1.5	1.0	1.3	1.0
16	4.3	4.0	5.3	1.3
15	7.3	7.3	10.0	7.4
14	12.3	12.5	9.3	8.4
13	9.5	7.8	6.3	10.4
12	12.5	5.0	5.8	11.4
11	6.5	8.3	7.0	6.0
10	9.8	8.5	7.0	8.0
9	9.3	6.8	9.3	9.4
8	6.5	9.5	7.3	8.7
7	4.8	6.5	6.8	7.0
6	3.3	5.5	6.5	4.0
≤ 5	2.5	2.5	3.3	2.0
Total[b]	90.4	85.2	85.2	85.0

[a] The equated delta was not computed for one item on each of five forms because fewer than 50% of the test takers reached that item. These forms were administered in the following months: December 1975, 1977, 1978, and 1984 and November 1980. For distributions with 84 items, each frequency was multiplied by 85/84 in computing the mean values.

[b] Because numbers ending with 5 in the hundredths place were rounded up, a given total does not necessarily equal the number of test items.

Table 16. Mean Numbers of Items by Item Difficulty (Equated Delta) for November and December SAT-Mathematical Test Forms Within Specified Periods from 1970 to 1984[a]

Equated Delta	*1970-1973*	*1974-1977*	*1978-1981*	*1982-1984*
		November Administrations		
≥ 18	1.3	2.5	1.5	2.7
17	3.8	3.0	3.8	4.3
16	4.8	4.5	5.3	4.3
15	5.3	4.5	4.3	4.3
14	4.0	4.8	3.8	4.3
13	4.3	3.5	3.3	3.7
12	6.3	6.3	6.1	4.7
11	9.1	7.8	8.3	7.3
10	7.1	5.8	7.1	9.7
9	6.8	6.8	8.1	5.3
8	3.3	3.8	3.3	6.0
7	2.5	4.5	2.0	1.7
6	1.0	1.8	2.3	1.0
≤ 5	0.5	0.8	0.8	0.7
Total[b]	60.1	60.4	60.2	60.0
		December Administrations		
≥ 18	1.3	1.3	1.3	1.4
17	2.5	2.5	3.5	3.0
16	5.0	6.6	4.8	5.7
15	4.5	3.3	4.0	6.1
14	6.3	4.8	4.0	2.4
13	5.8	4.3	6.8	4.0
12	5.5	5.3	5.0	5.7
11	7.3	5.8	8.8	5.8
10	6.5	8.8	7.0	8.8
9	8.0	8.6	6.3	6.4
8	3.8	3.5	3.3	5.1
7	1.8	2.8	2.0	3.0
6	1.3	2.0	2.0	1.3
≤ 5	0.8	0.5	1.3	1.4
Total[b]	60.4	60.1	60.1	60.1

[a] The equated delta was not computed for one item on each of five forms because fewer than 50% of the test takers reached that item. These forms were administered in the following months: December 1975, 1977, 1978, and 1984 and November 1980. Also, one item was not scored on the November 1981 form and on the December 1983 form. Therefore, the equated delta of these item were not included in the distributions. For distributions with 59 items, each frequency was multiplied by 60/59 in computing the mean values.

[b] Because numbers ending with 5 in the hundredths place were rounded up, a given total does not necessarily equal the number of test items.

BEST COPY AVAILABLE

Table 17. Scaled Scores Corresponding to Raw-Score Midpoints for November and December SAT Test Forms from 1970 to 1984

Year	*SAT-Verbal*	*SAT-Mathematical*
	November Administrations	
1970	517	544
1971	510	528
1972	493	533
1973	515	524
1974	480	536
1975	486	518
1976	486	522
1977	481	519
1978	480	525
1979	480	517
1980	461	523
1981	471	507
1982	471	507
1983	480	498
1984	478	518
	December Administrations	
1970	527	548
1971	526	534
1972	519	547
1973	508	539
1974	496	534
1975	486	528
1976	480	523
1977	495	527
1978	478	525
1979	481	533
1980	479	521
1981	474	538
1982	489	504
1983	466	507
1984	476	514

Table 18. Scaled-Score Ranges Corresponding to Selected Raw Scores for New SAT-Verbal Test Forms from March 1970 to January 1985

Raw Score[a]	March 1970 to April 1974 (20 Forms)	October 1974 to May 1978 (20 Forms)	October 1978 to December 1981 (24 Forms)	January 1982 to January 1985 (28 Forms)
90/85	800	8C0	800	800
85/80	760-800	740-770	750-770	720-750
79/75	720-760	710-730	700-730	670-710
74/70	690-720	670-700	660-690	630-670
69/65	650-690	640-660	610-650	600-640
64/60	620-650	600-620	580-620	570-600
58/55	580-610	570-590	540-580	540-570
53/50	540-580	530-550	510-550	510-540
48/45	510-550	500-520	480-510	470-500
42/40	480-510	460-480	440-470	440-470
37/35	440-470	430-450	410-440	410-440
32/30	410-440	390-410	380-410	380-410
26/25	370-400	350-370	350-370	340-380
21/20	340-370	320-340	310-340	310-340
16/15	310-340	280-310	280-310	270-310
11/10	270-310	250-270	240-270	240-270
5/ 5	230-270	210-240	210-240	200-240
0/ 0	200-240	200	200-210	200

[a] The raw scores on the left are for the 90-item forms administered from March 1970 to April 1974; those on the right are for the 85-item forms administered from October 1974 to January 198 .

Table 19. Scaled-Score Ranges Corresponding to Selected Raw Scores for New SAT-Mathematical Test Forms[a] from March 1970 to January 1985

Raw Score	March 1970 to April 1974 (20 Forms)	November 1974 to May 1978 (18 Forms)	October 1978 to December 1981 (21 Forms)	January 1982 to January 1985 (26 Forms)
60	800	800	800	800
55	750-800	740-770	740-760	720-750
50	700-750	690-720	690-710	670-710
45	660-700	640-670	640-670	630-670
40	610-650	600-630	600-630	590-620
35	570-600	560-580	560-580	540-570
30	520-560	520-540	510-540	490-530
25	460-510	470-490	470-490	440-490
20	440-460	430-450	430-450	400-450
15	390-410	390-410	380-410	360-400
10	340-370	350-370	340-360	320-360
5	300-330	300-330	290-320	280-310
0	250-290	260-290	250-280	240-280

[a] Forms with unscored items are not included.

Table 20. Mean Adjusted Proportions Correct and Observed Delta Statistics for November and December SAT Test Forms from 1970 to 1984

	SAT-Verbal			SAT-Mathematical		
Year	Mean Adj. Prop. Correct[a]	Mean Obs. Delta	SD Obs. Delta	Mean Adj. Prop. Correct[a]	Mean Obs. Delta	SD Obs. Delta
	November Administrations					
1970	.40	12.9	2.8	.41	12.9	2.7
1971	.41	12.8	2.7	.42	12.7	2.8
1972	.43	12.7	2.9	.42	12.8	2.8
1973	.40	12.9	2.8	.43	12.6	2.5
1974	.44	12.6	3.1	.40	12.8	2.8
1975	.42	12.7	3.0	.42	12.7	2.7
1976	.42	12.7	3.0	.41	12.8	2.6
1977	.43	12.7	3.0	.42	12.6	2.7
1978	.42	12.8	2.9	.40	12.9	3.0
1979	.42	12.7	3.1	.42	12.7	2.6
1980	.44	12.5	3.0	.41	12.6	2.8
1981	.43	12.7	3.0	.43	12.6	2.7
1982	.44	12.6	2.8	.43	12.5	2.8
1983	.43	12.7	2.8	.45	12.4	2.8
1984	.44	12.5	2.9	.44	12.5	2.9
	December Administrations					
1970	.38	13.2	2.6	.39	13.2	2.8
1971	.37	13.3	2.7	.40	13.0	2.9
1972	.36	13.3	2.7	.37	13.3	2.8
1973	.38	13.1	2.8	.36	13.4	2.7
1974	.38	13.2	3.0	.36	13.3	2.7
1975	.37	13.2	2.9	.36	13.2	2.6
1976	.38	13.2	2.9	.36	13.2	2.5
1977	.36	13.4	2.9	.34	13.4	2.5
1978	.36	13.4	3.0	.34	13.4	2.6
1979	.36	13.5	2.9	.33	13.6	2.6
1980	.38	13.2	2.8	.38	13.1	2.6
1981	.38	13.2	2.9	.34	13.4	2.5
1982	.37	13.2	2.7	.38	13.1	2.6
1983	.38	13.2	2.6	.40	12.9	2.5
1984	.39	13.1	2.6	.40	12.9	2.8

[a] Raw-score mean divided by the number of test items

Table 21. Percentages of Test Takers Completing 75% and 100% of Sections for November and December SAT-Verbal Test Forms from 1970 to 1984

	Verbal 1 (50/45 Items[a])		*Verbal 2 (40 Items)*	
Year	*75%*	*100%*	*75%*	*100%*
	November Administrations			
1970	99.5	77.8	99.7	71.6
1971	99.4	75.7	99.5	71.5
1972	99.6	76.4	99.6	77.1
1973	98.8	72.3	99.5	67.6
1974	99.6	88.0	95.6	70.9
1975	99.1	88.1	99.2	59.3
1976	94.5	76.0	99.8	82.6
1977	99.1	65.3	99.7	68.9
1978	98.8	74.3	99.6	74.1
1979	98.1	48.5	99.8	76.9
1980	95.2	44.4	98.2	79.0
1981	97.7	55.0	99.1	64.8
1982	95.7	51.1	99.9	74.6
1983	97.4	47.6	99.7	78.7
1984	97.0	65.3	99.8	65.5
	December Administrations			
1970	98.4	58.8	99.5	71.9
1971	98.2	52.7	99.8	68.8
1972	98.6	68.3	99.0	62.0
1973	99.3	69.7	99.7	76.1
1974	99.2	72.5	97.4	65.5
1975	98.1	49.8	98.4	62.8
1976	98.2	78.5	98.9	59.9
1977	96.8	57.5	98.7	68.8
1978	96.3	51.5	98.9	63.3
1979	94.5	65.0	99.7	70.7
1980	98.2	58.1	98.7	57.6
1981	95.6	47.2	99.7	70.0
1982	95.5	61.8	99.1	64.2
1983	95.3	42.4	99.3	73.8
1984	98.6	64.3	98.2	67.3

[a] From 1970 to 1973, 45 minutes were allowed for completion of this section. This time limit was reduced to 30 minutes for all forms administered from October 1974 on.

Table 22. Percentages of Test Takers Completing 75% and 100% of Sections for November and December SAT-Mathematical Test Forms from 1970 to 1984

Year	*Mathematical 1 (25 Items)*		*Mathematical 2 (35 Items[a])*	
	75%	*100%*	*75%*	*100%*
		November Administrations		
1970	95.2	43.2	99.4	94.1
1971	96.7	38.3	99.1	77.5
1972	97.4	60.5	99.5	92.5
1973	97.0	53.3	98.5	85.5
1974	99.1	63.1	92.0	66.5
1975	98.1	81.3	98.7	72.0
1976	96.6	68.2	99.5	83.1
1977	99.2	76.6	98.5	54.5
1978	99.8	77.0	97.5	54.9
1979	98.9	68.1	98.3	54.2
1980	98.9	84.6	98.5	46.2
1981	99.1	65.4	99.4	72.9
1982	99.4	62.1	98.7	58.0
1983	99.3	66.4	99.7	64.8
1984	99.5	61.4	99.6	72.8
		December Administrations		
1970	95.0	66.9	98.3	85.2
1971	95.7	86.0	98.4	86.3
1972	97.1	68.9	96.1	64.3
1973	95.9	69.5	96.6	77.1
1974	97.9	68.7	94.7	75.8
1975	97.6	56.7	97.1	45.0
1976	98.3	76.8	98.6	54.2
1977	98.2	71.3	98.7	48.3
1978	97.9	55.0	99.1	38.5
1979	98.6	81.1	98.8	53.2
1980	98.6	77.8	99.2	51.0
1981	95.5	64.8	98.5	52.4
1982	98.3	64.0	99.0	61.4
1983	97.4	67.6	99.6	56.4
1984	98.6	86.6	98.9	49.1

[a] From 1970 to 1973, 45 minutes were allowed for completion of this section. This time limit was reduced to 30 minutes for all forms administered from October 1974 on.

Table 23. Means and Standard Deviations of the Numbers of Items Not Reached and Ratios of Variances for Sections of November and December SAT-Verbal Test Forms from 1970 to 1984

	Verbal 1 (50/45 Items[a])			Verbal 2 (40-Items)		
Year	Mean	SD	Ratio of Variances[b]	Mean	SD	Ratio of Variances[b]
			November Administrations			
1970	1.05	2.58	.07	.82	1.75	.05
1971	1.09	2.47	.07	1.18	2.24	.07
1972	.91	2.25	.06	.88	2.00	.06
1973	1.36	2.85	.08	1.30	2.35	.09
1974	.53	2.24	.08	1.57	3.50	.21
1975	.56	2.25	.07	1.86	2.83	.13
1976	1.75	4.02	.20	.80	2.09	.07
1977	.85	2.01	.06	1.34	2.45	.10
1978	.95	2.29	.07	1.05	2.15	.08
1979	1.72	2.82	.12	.87	1.94	.07
1980	2.14	3.75	.18	.78	1.93	.06
1981	1.48	2.96	.10	1.27	2.40	.12
1982	2.13	3.79	.18	.95	1.96	.06
1983	1.88	3.15	.12	.72	1.78	.05
1984	1.74	3.38	.16	1.37	2.27	.07
			December Administrations			
1970	2.10	3.36	.13	1.02	2.17	.07
1971	2.61	3.76	.18	1.11	2.11	.07
1972	1.61	3.19	.14	1.72	2.74	.11
1973	1.65	3.17	.11	.91	2.02	.08
1974	1.29	2.67	.11	1.04	2.71	.12
1975	1.36	2.79	.20	1.84	3.13	.14
1976	1.09	2.83	.12	1.60	2.67	.12
1977	1.67	3.37	.17	1.35	2.52	.10
1978	2.27	3.68	.21	1.53	2.70	.12
1979	2.26	4.15	.24	1.17	2.28	.10
1980	1.80	2.90	.13	1.83	2.80	.12
1981	2.16	3.82	.20	1.14	2.16	.08
1982	2.05	3.79	.18	1.14	2.25	.10
1983	2.53	3.84	.18	1.08	2.33	.08
1984	1.33	2.54	.09	1.71	3.02	.13

[a] From 1970 to 1973, 45 minutes were allowed for completion of this section. This time limit was reduced to 30 minutes for all forms administered from October 1974 on.

[b] The variance of not reached items in the section divided by the variance of the total score on the section.

Table 24. Means and Standard Deviations of the Numbers of Items Not Reached and Ratios of Variances for Sections of November and December SAT-Mathematical Test Forms from 1970 to 1984

	Mathematical 1 (25 Items)			*Mathematical 2 (35 Items[a])*		
Year	*Mean*	*SD*	*Ratio of Variances[b]*	*Mean*	*SD*	*Ratio of Variances[b]*
			November Administrations			
1970	1.70	2.34	.18	.24	1.41	.04
1971	1.97	2.34	.19	.64	1.76	.05
1972	.96	1.81	.10	.28	1.39	.03
1973	1.06	1.87	.10	.45	1.86	.06
1974	.65	1.24	.05	1.96	3.93	.28
1975	.57	1.55	.07	1.20	2.48	.08
1976	1.06	2.00	.11	.60	1.73	.04
1977	.48	1.18	.04	1.66	2.67	.12
1978	.46	1.15	.04	2.09	3.09	.19
1979	.69	1.53	.06	1.86	2.71	.12
1980	.49	1.41	.06	1.57	2.40	.10
1981	.61	1.28	.05	.95	1.99	.06
1982	.72	1.29	.05	1.12	2.06	.07
1983	.80	1.48	.07	1.03	1.88	.06
1984	.69	1.28	.05	.71	1.55	.04
			December Administrations			
1970	1.25	2.38	.21	.58	2.16	.08
1971	1.24	2.31	.17	.55	2.20	.10
1972	1.07	2.15	.15	1.02	3.06	.17
1973	1.24	2.43	.19	1.00	2.60	.14
1974	.91	1.75	.10	1.31	3.31	.18
1975	1.33	2.07	.12	2.37	3.27	.17
1976	.64	1.80	.06	1.45	2.37	.08
1977	.62	1.47	.06	1.23	2.19	.08
1978	.96	1.73	.09	1.62	2.14	.08
1979	.84	1.71	.10	1.46	2.30	.08
1980	.59	1.49	.06	1.52	2.19	.08
1981	1.28	2.32	.14	1.58	2.39	.10
1982	.95	1.69	.09	1.02	2.09	.06
1983	1.04	1.96	.12	1.06	1.78	.05
1984	.44	1.45	.06	1.81	2.46	.09

[a] From 1970 to 1973, 45 minutes were allowed for completion of this section. This time limit was reduced to 30 minutes for all forms administered from October 1974 on.

[b] The variance of not-reached items in the section divided by the variance of the total score on the section.

BEST COPY AVAILABLE

Table 25. Reliability Coefficients and Standard Errors of Measurement for November and December SAT Test Forms from 1970 to 1984

	SAT-Verbal			SAT-Mathematical		
Year	Reliability	Rel. (SD=100)[a]	Standard Error of Measurement	Reliability	Rel. (SD=100)[a]	Standard Error of Measurement
			November Administrations			
1970	.915	.906	30.7	.902	.885	33.9
1971	.918	.901	31.0	.907	.882	34.4
1972	.921	.911	29.8	.913	.888	33.5
1973	.922	.909	30.1	.913	.892	32.8
1974	.909	.896	32.2	.906	.882	34.3
1975	.917	.906	30.6	.925	.892	32.9
1976	.924	.910	30.0	.922	.894	32.6
1977	.909	.901	31.5	.912	.890	33.2
1978	.913	.907	30.5	.898	.875	35.4
1979	.908	.902	31.3	.912	.886	33.7
1980	.921	.912	29.7	.911	.882	34.3
1981	.920	.907	30.5	.914	.891	33.0
1982	.926	.914	29.3	.911	.881	34.5
1983	.926	.902	31.3	.913	.883	34.2
1984	.922	.900	31.7	.912	.878	34.9
			December Administrations			
1970	.915	.895	32.4	.905	.875	35.3
1971	.915	.901	31.4	.903	.884	34.1
1972	.910	.901	31.5	.908	.886	33.7
1973	.908	.887	33.6	.910	.886	33.8
1974	.909	.898	32.0	.903	.876	35.2
1975	.920	.903	31.2	.916	.892	32.9
1976	.910	.899	31.8	.922	.905	30.8
1977	.914	.901	31.5	.916	.890	33.2
1978	.912	.905	30.9	.904	.883	34.2
1979	.907	.897	32.1	.908	.882	34.3
1980	.910	.906	30.7	.909	.883	34.2
1981	.916	.911	29.8	.910	.880	34.6
1982	.911	.908	30.4	.914	.893	32.7
1983	.921	.905	30.8	.915	.888	33.4
1984	.916	.900	31.7	.925	.890	33.1

[a] Reliabilities were estimated for a hypothetical reference group with a scaled-score standard deviation of 100.

Table 26. Test-Retest Correlations for the SAT from 1970 to 1984[a,b,c]

Year	March/April-November[d]	May-November	June/July-November	March/April-December[d]	May-December	June/July-December
			SAT-Verbal			
1970	.90	.89	.87	.90	.89	.88
1971	.89		.87	.89		.88
1972	.89		.88	.89		.88
1973	.89		.88	.87		.88
1974	.88		.88	.87		.88
1975	.88		.87	.87		.87
1976	.88		.88	.87		.88
1977	.88	.88	.88	.88	.88	.88
1978	.88	.88	.87	.87	.88	.88
1979	.89	.88	.87	.87	.88	.86
1980	.89	.88	.88	.87	.87	.87
1981	.88	.88	.88	.88	.87	.88
1982	.89	.88	.88	.89	.86	.88
1983	.89	.88	.88	.89	.89	.88
1984	.89	.89	.89	.88	.89	.88
			SAT-Mathematical			
1970	.88	.86	.86	.88	.87	.87
1971	.88		.86	.88		.87
1972	.88		.87	.89		.88
1973	.88		.87	.88		.89
1974	.87		.87	.87		.88
1975	.89		.87	.88		.88
1976	.89		.88	.90		.90
1977	.88	.89	.88	.88	.89	.88
1978	.88	.88	.88	.88	.88	.88
1979	.88	.88	.88	.88	.88	.88
1980	.88	.87	.86	.87	.87	.87
1981	.87	.88	.88	.87	.88	.88
1982	.88	.87	.88	.89	.88	.88
1983	.88	.88	.89	.88	.89	.88
1984	.88	.87	.87	.89	.88	.88

[a] Aceptation of Table 3.9 in Donlon (1984)

[b] These correletione are based on students who took the SAT in the spring of their junior year in secondery school and repeated the test in the fall of their senior year.

[c] The junior-year to senior-year testing patterns with the largest numbers of repeaters were the following:

Year(s)	Senior-Year Administration	Junior-Year Administration	Sample Size Range (in 000's)
1970	November	May	126 - 126
1971-77	November	March/April	77 - 165
1978-84	November	May	81 - 117
1970	December	May	66 - 66
1971-76	December	March/April	39 - 56
1977-84	December	May	21 - 29

[d] Data are from the March administration in 1970, the April administrations from 1971 to 1976, and the March administrations from 1977 to 1984.

Table 27. Correlations Among SAT-V, SAT-M, and TSWE, Including Correlations Corrected for Attenuation, for November and December SAT Test Forms from 1970 to 1984

	SAT-V and SAT-M		*SAT-V and TSWE*		*SAT-M and TSWE*	
Year	*Original Correlation*	*Corrected Correlation*	*Original Correlation*	*Corrected Correlation*	*Original Correlation*	*Corrected Correlation*
			November Administrations			
1970	.88	.75				
1971	.68	.75				
1972	.71	.77				
1973	.67	.73				
1974	.67	.74	.75	.83	.59	.66
1975	.68	.74	.77	.85	.62	.68
1976	.67	.73	.79	.88	.63	.70
1977	.67	.73	.76	.85	.62	.69
1978	.67	.74	.77	.85	.61	.69
1979	.84	.70	.79	.88	.62	.69
1980	.70	.77	.78	.86	.63	.70
1981	.87	.72	.80	.88	.64	.71
1982	.86	.72	.79	.86	.64	.70
1983	.86	.72	.77	.85	.61	.69
1984	.86	.72	.78	.84	.65	.71
			December Administrations			
1970	.88	.75				
1971	.86	.72				
1972	.69	.75				
1973	.89	.76				
1974	.62	.69	.78	.87	.59	.86
1975	.68	.74	.78	.86	.62	.89
1976	.85	.70	.78	.87	.63	.70
1977	.87	.73	.79	.88	.65	.72
1978	.84	.71	.81	.90	.63	.70
1979	.88	.75	.80	.89	.63	.70
1980	.87	.74	.79	.88	.62	.69
1981	.63	.69	.78	.86	.55	.61
1982	.88	.74	.78	.88	.65	.72
1983	.66	.72	.79	.88	.62	.69
1984	.64	.70	.77	.87	.60	.67

Table 28. Correlations of Verbal Sections and of Mathematical Sections, Including Correlations Corrected for Attenuation, for November and December SAT Test Forms from 1970 to 1984

	SAT-Verbal Sections		*SAT-Mathematical Sections*	
Year	*Original Correlation*	*Corrected Correlation*	*Original Correlation*	*Corrected Correlation*
		November Administrations		
1970	.84	1.00	.80	.98
1971	.83	.98	.81	.98
1972	.84	.99	.82	.98
1973	.84	.98	.84	1.00
1974	.82	.98	.82	.99
1975	.83	.97	.84	.98
1976	.84	.98	.85	1.00
1977	.82	.98	.83	.99
1978	.83	.99	.81	.99
1979	.83	1.00	.83	.98
1980	.83	.97	.83	.99
1981	.84	.99	.82	.98
1982	.85	.99	.83	1.00
1983	.84	.98	.83	.99
1984	.83	.97	.84	1.00
		December Administrations		
1970	.83	.98	.81	.99
1971	.83	.98	.81	.98
1972	.82	.97	.83	.99
1973	.81	.97	.81	.97
1974	.81	.97	.82	1.00
1975	.84	.98	.83	.98
1976	.83	1.00	.85	.99
1977	.83	.99	.85	1.00
1978	.83	.99	.82	1.00
1979	.81	.97	.81	.98
1980	.81	.97	.83	.99
1981	.82	.97	.83	.99
1982	.82	.99	.83	.98
1983	.84	.98	.83	.98
1984	.84	.99	.86	1.00

Table 29. Correlations Between SAT Reading and Vocabulary Subscores, Including Correlations Corrected for Attenuation, for November and December SAT Test Forms from 1974 to 1984

Year	*Original Correlation*	*Corrected Correlation*
	November Administrations	
1974	.78	.93
1975	.80	.94
1976	.80	.92
1977	.80	.95
1978	.81	.96
1979	.79	.94
1980	.81	.95
1981	.81	.94
1982	.80	.92
1983	.81	.94
1984	.80	.93
	December Administrations	
1974	.78	.93
1975	.81	.95
1976	.78	.93
1977	.81	.95
1978	.80	.95
1979	.78	.94
1980	.78	.94
1981	.80	.94
1982	.80	.95
1983	.80	.94
1984	.80	.94

Table 30a. Standardized Slopes and Differences[a] for SAT-V Forms Administered in November and December from 1970 to 1984

Variable	Mean	SD	Overall Slope	Slopes Within Periods		Difference Between Period Means
			1970-1984	1970-1973	1974-1984	1970-73 vs. 1974-84
			Test Difficulty			
Mean Equated Delta	11.46	.21	-.18	-.35	-.06	-1.91
Std. Dev. of Equated Delta	3.11	.19	.04	.50	-.17	1.21
Scaled Score Corresponding to the Raw Score Midpoint	488	18	-.20	-.24	-.10	-1.94
			Item-Test Correlations			
Mean Biserial Correlation	.47	.02	.13	.00	.12	1.01
			Relative Test Difficulty			
Mean Observed Delta	12.97	.30	-.03	-.07	-.03	-.23
Raw Score Mean/ No. of Items	.40	.03	.04	.02	.04	.37
			Speededness			
Section 1						
Percentage of Test Takers Completing 75% of the Test	97.67	1.63	-.13	.04	-.13	-1.09
Ratio of Not-Reached and Total-Test Variance	.13	.05	.08	-.08	.06	.74
Mean No. of Not-Reached Items	1.60	.56	-.08	.14	.19	.12
Section 2						
Percentage of Test Takers Completing 75% of the Test	99.16	.88	.02	-.04	.16	-.59
Ratio of Not-Reached and Total-Test Variance	.09	.03	-.00(4)	.26	-.18	.78
Mean No. of Not-Reached Items	1.23	.34	-.00(4)	-.21	-.09	.45
			Reliability			
Internal-Consistency Reliability	.916	.006	.05	-.04	.14	-.02
Scaled-Score Standard Error of Measurement	31.07	.94	-.08	.04	-.08	-.36
Adjusted Internal-Consistency Reliability	.903	.006	.06	-.07	.08	.40
Test-Retest Correlation	.88	.009	-.03	-.66	.16	-1.04

[a] The slopes and differences for each variable were divided by the standard deviations for the variable.

Table 30b. Standardized Slopes and Differences[a] for SAT-M Forms Administered in November and December from 1970 to 1984

Variable	Mean	SD	Overall Slope	Slopes Within Periods		Difference Between Period Means
			1970-1984	1970-1973	1974-1984	1970-73 vs. 1974-84
			Test Difficulty			
Mean Equated Delta	12.23	.20	-.03	.43	.09	-.94
Std. Dev. of Equated Delta	3.19	.21	.09	.56	.04	.83
Scaled Score Corresponding to the Raw Score Midpoint	524	12	-.17	-.27	-.16	-1.34
			Item-Test Correlations			
Mean Biserial Correlation	.54	.02	.07	.37	.05	.51
			Relative Test Difficulty			
Mean Observed Delta	12.96	.32	-.05	.02	-.11	-.10
Raw Score Mean/ No. of Items	.39	.03	.03	-.09	.12	-.25
			Speededness			
Section 1						
Percentage of Test Takers Completing 75% of the Test	97.83	1.37	.15	.37	.04	1.57
Ratio of Not-Reached and Total-Test Variance	.10	.05	-.15	-.41	-.01	-1.72
Mean No. of Not-Reached Items	.90	.37	.13	.42	-.02	-1.49
Section 2						
Percentage of Test Takers Completing 75% of the Test	98.28	1.63	.08	-.30	.21	.04
Ratio of Not-Reached and Total-Test Variance	.10	.06	-.05	.26	-.19	.30
Mean No. of Not-Reachod Items	1.20	.55	-.09	-.18	-.09	1.50
			Reliability			
Internal-Consistency Reliability	.911	.007	.08	.45	.03	.76
Scaled-Score Standard Error of Measurement	33.77	.97	- 00(4)	-.47	.06	-.24
Adjusted Internal-Consistency Reliability	.886	.007	.00(2)	.48	-.06	.22
Test-Retest Correlation	.88	.007	-.01	.07	.00	-.12

[a] The slopes and differences for each variable were divided by the standard deviations for the variable.

Table 30c. Standardized Slopes and Differences[a] for SAT-V, SAT-M, and TSWE Forms Administered in November and December from 1970 to 1984

Variable	Mean	SD	Overall Slope	Slopes Within Periods		Difference Between Period Means
			1970-1984	1970-1973	1974-1984	1970-73 vs. 1974-84
			Raw Correlations			
SAT-V & SAT-M	.668	.019	-.08	.08	-.01	-.91
SAT-V & TSWE	.781	.014	N/A	N/A	.05	N/A
SAT-M & TSWE	.620	.022	N/A	N/A	.04	N/A
SAT-V Sections 1 & 2	.828	.012	.04	-.25	.10	.05
SAT-M Sections 1 & 2	.826	.014	.10	.47	.02	1.05
SAT Reading & Vocabulary	.797	.011	N/A	N/A	.12	N/A
			Corrected for Attenuation			
SAT-V & SAT-M	.731	.021	-.09	.04	-.03	-1.02
SAT-V & TSWE	.866	.016	N/A	N/A	.04	N/A
SAT-M & TSWE	.690	.024	N/A	N/A	.04	N/A
SAT-V Sections 1 & 2	.980	.009	N/A	-.30	-.02	.11
SAT-M Sections 1 & 2	.988	.008	.09	-.00(2)	.02	1.05
SAT Reading & Vocabulary	.940	.010	N/A	N/A	.00(3)	N/A

[a] The regression weights for the slopes and differences for each variable were divided by the standard deviations for the variable.

[b] The Test of Standard Written English and the SAT Reading and Vocabulary Subscores were introduced in 1974.

Table 31. Summary of Changes in SAT Statistical Specifications from March 1970 to January 1985

Beginning Date	*Change*
October 1974	SAT-V statistical specifications changed: o Mean item difficulty (delta) reduced from 11.7 to 11.4 o Standard deviation of item difficulties (deltas) increased from 2.9 to 3.3 o Number of difficult (delta = 15 and above) items increased from 14 (out of 90) to 16 (out of 85); number of easy (delta = 8 and below) items increased from 18 (out of 90) to 25 (out of 85) o Mean biserial item-total correlation increased from .42 to .43 (in terms of pretest statistics)
	SAT-M statistical specifications changed: o Mean item difficulty (delta) reduced from 12.5 to 12.2 o Standard deviation of item difficulties (deltas) increased from 3.1 to 3.2 o Number of difficult (delta = 15 and above) items remained at 15; number of easy (delta = 8 and below) items increased from 7 to 9 o Mean biserial item-total correlation remained at .47 (in terms of pretest statistics)
January 1982	Number of difficult SAT-V items reduced and statistical specifications changed: o Standard deviation of item difficulties (deltas) reduced from 3.3 to 3.0 o Number of difficult (delta = 15 and above) items reduced from 16 to 8; number of moderately difficult (delta = 13-14) items increased from 16 to 24

Table 32. Reliability and Validity as a Function of Reduced Test Length

Original Reliability	*New Reliability*	*New Validity Corresponding to Original Validity of:* .30	.33	.36	.39	.42
Test Shortened by 1/18th						
.94	.94	.30	.33	.36	.39	.42
.92	.92	.30	.33	.36	.39	.42
.90	.89	.30	.33	.36	.39	.42
.88	.87	.30	.33	.36	.39	.42
.86	.85	.30	.33	.35	.39	.42
.84	.83	.30	.33	.36	.39	.42
Test Shortened by 1/5th						
.94	.93	.30	.33	.36	.39	.42
.92	.90	.30	.33	.36	.39	.42
.90	.88	.30	.33	.36	.39	.41
.88	.85	.30	.33	.35	.38	.41
.86	.83	.29	.32	.35	.38	.41
.84	.81	.29	.32	.35	.38	.41

Table 33. Equating Methods[a] Used for November and December SAT-Verbal and SAT-Mathematical Equatings from 1970 to 1984

	SAT-Verbal		SAT-Mathematical	
Year	First Equating	Second Equating	First Equating	Second Equating
	November Administrations			
1970	Tucker	Tucker	Tucker	Tucker
1971	Tucker	Tucker	Tucker	Tucker
1972	Tucker	Tucker	Tucker	Tucker
1973	Tucker	Tucker	Tucker	Tucker
1974	Tucker	Tucker	Tucker	Tucker
1975	Tucker	Tucker	Tucker	Tucker
1976	Tucker	Tucker	Tucker	Tucker
1977	Tucker	Tucker	Tucker	Tucker
1978	Tucker	Tucker	Tucker	Tucker
1979	Tucker	Levine	Tucker	Tucker
1980	Tucker	Tucker	Tucker	Tucker
1981	Tucker	Tucker	Tucker	Tucker
1982	IRT	IRT	IRT	IRT
1983	IRT	IRT	IRT	IRT
1984	IRT	IRT	IRT/Tucker[b]	Tucker/Tucker[b]
	December Administrations			
1970	Tucker	Tucker	Tucker	Tucker
1971	Tucker	Tucker	Tucker	Tucker
1972	Levine	Tucker	Tucker	Tucker
1973	Tucker	Levine	Tucker	Levine
1974	Tucker	Tucker	Tucker	Tucker
1975	Levine	Levine	Levine	Tucker
1976	Tucker	Tucker	Tucker	Tucker
1977	Tucker	Levine	Tucker	Levine
1978	Levine	Tucker	Levine	Tucker
1979	Tucker	Tucker	Tucker	Tucker
1980	Tucker	Tucker	Tucker	Tucker
1981	Tucker	Levine	Tucker	Levine
1982	IRT	IRT	IRT	IRT
1983	IRT	Levine	IRT	IRT
1984	IRT/IRT[b]	Tucker/Tucker[b]	IRT/IRT[b]	Tucker/Tucker[b]

[a] "IRT" refers to item response theory.

[b] This equating went back to the two parent forms for the old form rather than to the old form itself. One of the parent forms for the old form used in this equating was the same as one of the parent forms used in the other equating. Therefore, in averaging the two equating lines, the equating to the common parent form was weighted half as much as the equating to the distinct parent form.

Table 34. New- and Old-Form Equating Sample Sizes, Scaled-Score Means, and Scaled-Score Standard Deviations for November and December SAT-Verbal Test Forms from 1970 to 1984[a]

Year	First Equating						Second Equating					
	New Form			Old Form			New Form			Old Form		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
						November Administrations						
1970	5000	457	107	4866	450	105	5000	456	107	5000	465	110
1971	5000	458	107	2565	464	110	5000	460	108	5009	471	111
1972	5000	451	105	5000	462	107	5000	448	106	5000	457	107
1973	5084	457	108	5000	470	106	5084	460	106	5000	454	108
1974	5127	443	108	4820	441	108	5280	446	109	5000	457	107
1975	5859	443	107	4237	424	108	5787	439	107	5000	456	106
1976	5775	440	104	4920	441	108	5688	441	105	5581	438	105
1977	6232	431	107	6005	429	109	6213	434	108	4929	440	109
1978	7969	431	106	5161	419	107	7839	433	105	5581	438	105
1979	6368	434	108	7623	431	107	6327	435	106	5364	407	106
1980	6350	430	106	6327	435	106	6355	428	106	7406	408	110
1981	7394	432	107	5162	433	106	7395	430	106	4951	424	111
1982	8245	434	103	5693	431	107	8290	436	105	4788	400	108
1983	5560	436	102	5173	435	105	5513	436	105	4769	427	110
1984	5530	440	107	4977	437	107	5214	441	107	4966	425	110
						December Administrations						
1970	4622	462	107	4996	457	108	4584	459	109	5023	464	112
1971	5000	445	107	5000	461	102	5000	448	108	4622	462	112
1972	4845	440	105	5000	472	107	4845	438	106	5000	450	107
1973	5000	431	108	5000	457	107	5000	438	110	4885	478	113
1974	6276	429	108	5100	457	111	6182	424	106	5303	420	108
1975	5348	408	107	5280	446	109	4980	408	110	4814	441	108
1976	5668	412	104	5000	436	109	5634	410	106	5196	411	108
1977	4951	414	107	4711	412	106	4916	411	107	5000	431	110
1978	5453	400	106	5462	434	107	6416	403	103	5196	411	108
1979	6284	397	108	6173	399	103	6197	397	104	6949	405	111
1980	7834	410	106	6284	397	104	7751	407	103	4793	433	106
1981	4936	403	107	4729	406	102	5002	403	101	5072	432	104
1982	5503	403	103	4639	406	111	5433	403	102	8777	420	108
1983	6536	402	103	6185	403	104	6450	400	103	4356	431	105
1984	6872	407	102	4328	401	104	6785	407	101	4482	437	105

[a] The means and standard deviations for the years 1982-84 are estimates from linear conversions that were not used to report scores.

Table 35. New- and Old-Form Equating Sample Sizes, Scaled-Score Means, and Scaled-Score Standard Deviations for November and December SAT-Mathematical Test Forms from 1970 to 1984[a]

	First Equating						Second Equating					
	New Form			Old Form			New Form			Old Form		
Year	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
November Administrations												
1970	5000	494	109	4371	489	111	5000	492	108	5000	502	113
1971	5000	488	111	4494	491	116	5000	488	112	5000	503	112
1972	5000	486	111	5000	492	111	5000	486	111	5000	494	109
1973	5084	489	112	5000	499	113	5084	490	113	5000	486	109
1974	5032	483	112	4063	476	113	5177	480	110	5000	494	110
1975	5732	480	115	4234	462	115	5608	480	115	5000	488	109
1976	5673	480	117	4849	470	113	5802	478	117	5511	483	118
1977	6183	468	113	5833	471	115	6043	471	115	4867	472	113
1978	7792	469	113	5022	465	110	7675	471	114	5511	483	118
1979	6261	478	114	7566	470	113	6164	474	114	5281	457	114
1980	6296	473	114	6164	474	114	6183	470	114	7190	450	118
1981	7299	469	110	5086	475	113	7158	470	112	4898	456	112
1982	8155	472	110	5542	471	113	8048	473	112	4706	443	117
1983	5428	476	111	5021	473	112	5333	475	110	4707	449	99
1984	5149	484	114	4813	478	111	5126	485	112	5036	468	114
December Administrations												
1970	4631	491	115	5000	493	116	4668	491	117	5022	494	117
1971	5000	486	116	5000	498	112	5000	483	113	5000	481	112
1972	4845	483	112	5000	493	106	4845	479	113	4631	491	115
1973	5000	464	113	5000	488	110	5000	470	115	5000	505	108
1974	5990	463	113	4442	455	112	6098	461	112	5094	491	114
1975	5275	449	115	5032	483	112	5051	452	114	4823	476	114
1976	5554	455	113	5000	464	113	5474	453	114	5128	456	104
1977	4828	444	114	4548	443	117	4829	442	110	5000	458	113
1978	6323	443	112	5369	480	111	6232	445	110	5272	449	115
1979	6103	444	111	6087	443	110	5984	446	113	6880	452	117
1980	7654	447	111	6103	444	111	7530	452	113	4709	469	115
1981	4853	450	112	4647	454	113	4780	446	112	4972	476	112
1982	5405	447	111	4552	450	115	5281	450	112	8752	457	112
1983	6286	451	113	6063	447	111	8212	449	114	4299	480	114
1984	6668	461	110	4316	452	113	6536	459	109	4406	484	113

[a] The means and standard deviations for the years 1982-84 are estimates from linear conversions that were not used to report scores.

Table 36. Comparisons of Operational and Equipercentile Equating Lines at the Midpoints of the Raw-Score Ranges for November and December SAT Test Forms from 1970 to 1984

	SAT-Verbal			SAT-Mathematical		
Year	Dates of Old Form Administrations	Operational Conversion at Midpoint	Oper. Line Minus Equi. Line	Dates of Old Form Administrations	Operational Conversion at Midpoint	Oper. Line Minus Equi. Line
			November Administrations			
1970	1-2/67; 11/69	517.3	3.8	1-2/67; 11/69	543.5	.5
1971	1-2/66; 11/69	509.5	.5	3-4/67; 11/69	527.9	4.4
1972	3-4/68; 11/70	494.0	6.0	3-4/68; 11/70	532.7	4.2
1973	12/68; 4-5/71	514.7	2.2	12/68; 4-5/71	524.4	1.9
1974	11/70; 3-4/72	481.0	.0	11/70; 3-4/72	535.7	-2.8
1975	4-5/71; 1-2/74	486.3	-1.4	4-5/71; 1-2/74	518.2	6.2
1976	3-4/73; 11/75	486.3	1.5	3-4/73; 11/75	521.6	- .6
1977	3-4/73; 1-2/76	480.7	3.0	3-4/73; 1-2/76	518.5	3.6
1978	11/75; 3-4/77	480.2	.4	11/75; 3-4/77	524.8	-2.4
1979	12/76; 11/78	480.3	.0	12/76; 11/78	517.4	-2.7
1980	1-2/78; 11/79	461.4	4.8	1-2/78; 11/79	523.2	- .2
1981	3-4/79; 11/80	470.9	1.9	3-4/79; 11/80	506.8	2.3
1982	1-2/80; 11/81	470.7	.5	1-2/80; 11/81	506.7	3.8
1983	3-4/81; 11/82	479.9	3.5	3-4/81; 11/82	498.0	-2.2
1984	3-4/82; 11/83	477.7	-5.7	3-4/82; 11/83	518.2	1.2
			December Administrations			
1970	3-4/67; 12/69	526.6	- .4	3-4/67; 12/69	549.0	1.0
1971	4-5/68; 12/70	527.0	.5	12/67; 3-4/70	534.0	3.0
1972	12/67; 3-4/70	518.5	.5	4-5/68; 12/70	547.2	- .3
1973	11/68; 4-5/71	507.9	-3.1	11/68; 4-5/71	539.1	-2.4
1974	12/70; 1-2/73	496.4	3.4	12/70; 1-2/73	533.7	2.2
1975	3-4/72; 11/74	486.4	.9	3-4/72; 11/74	527.7	5.2
1976	12/73; 12/75	480.4	3.4	12/73; 12/75	522.8	7.3
1977	12/73; 3-4/76	496.0	- .6	12/73; 3-4/76	527.3	2.8
1978	12/75; 4-5/77	478.0	-1.8	12/75; 4-5/77	525.3	-6.6
1979	1-2/77; 12/78	481.2	5.3	1-2/77; 12/78	533.3	1.3
1980	4-5/78; 12/79	478.6	6.3	4-5/78; 12/79	521.0	1.0
1981	4-5/79; 12/80	473.8	.8	4-5/79; 12/80	538.5	- .7
1982	3-4/80; 12/81	469.2	-3.3	3-4/80; 12/81	503.9	4.3
1983	4-5/81; 12/82	466.1	.1	4-5/81; 12/82	506.8	-5.5
1984	4-5/82; 12/83	475.5	-2.0	4-5/82; 12/83	513.6	1.9

Note: Linear equating was used operationally from 1970 to 1984, and curvilinear equating, sometimes in combination with linear equating, from 1982 to 1984.

BEST COPY AVAILABLE

Table 37. Comparisons of Operational and Equipercentile Equating Lines at Midpoints of the Raw-Score Ranges Between and Within Periods

Period	Type of Comparison	SAT-Verbal Mean	SAT-Verbal Range	SAT-Mathematical Mean	SAT-Mathematical Range
		November Administrations			
1970-73	Within	3.13	.5 - 6.0	2.75	.5 - 4.4
1978-81	Within	1.78	.0 - 4.8	1.91	.2 - 2.7
1984	Within	5.71	5.7 - 5.7	1.15	1.2 - 1.2
	Within	2.81	.0 - 6.0	2.20	.2 - 4.4
1976-77	Mixed	2.25	1.5 - 3.0	2.10	.6 - 3.6
1983	Mixed	3.46	3.5 - 3.5	2.24	2.2 - 2.2
	Mixed	2.65	1.5 - 3.5	2.15	.6 - 3.6
1974-75	Between	.71	.0 - 1.4	4.50	2.8 - 6.2
1982	Between	.49	.5 - .5	3.83	3.8 - 3.8
	Between	.63	.0 - 1.4	4.28	2.8 - 6.2
1970-84	All	2.34	.0 - 6.0	2.60	.2 - 6.2
		December Administrations			
1970-73	Within	1.13	.4 - 3.1	1.68	.3 - 3.0
1978-81	Within	3.54	.8 - 6.3	2.39	.7 - 6.6
1984	Within	1.97	2.0 - 2.0	1.91	1.9 - 1.9
	Within	2.29	.4 - 6.3	2.02	.3 - 6.6
1975-77	Mixed	1.63	.6 - 3.4	5.10	2.8 - 7.3
1983	Mixed	.09	.1 - .1	5.49	5.5 - 5.5
	Mixed	1.25	.1 - 3.4	5.20	2.8 - 7.3
1974	Between	3.40	3.4 - 3.4	2.20	2.2 - 2.2
1982	Between	3.32	3.3 - 3.3	4.33	4.3 - 4.3
	Between	3.36	3.3 - 3.4	3.27	2.2 - 4.3
1970-84	All	2.16	.1 - 6.3	3.03	.3 - 7.3

Table 38. Comparisons of Operational and Experimental Conversions for Selected SAT-Verbal Test Forms

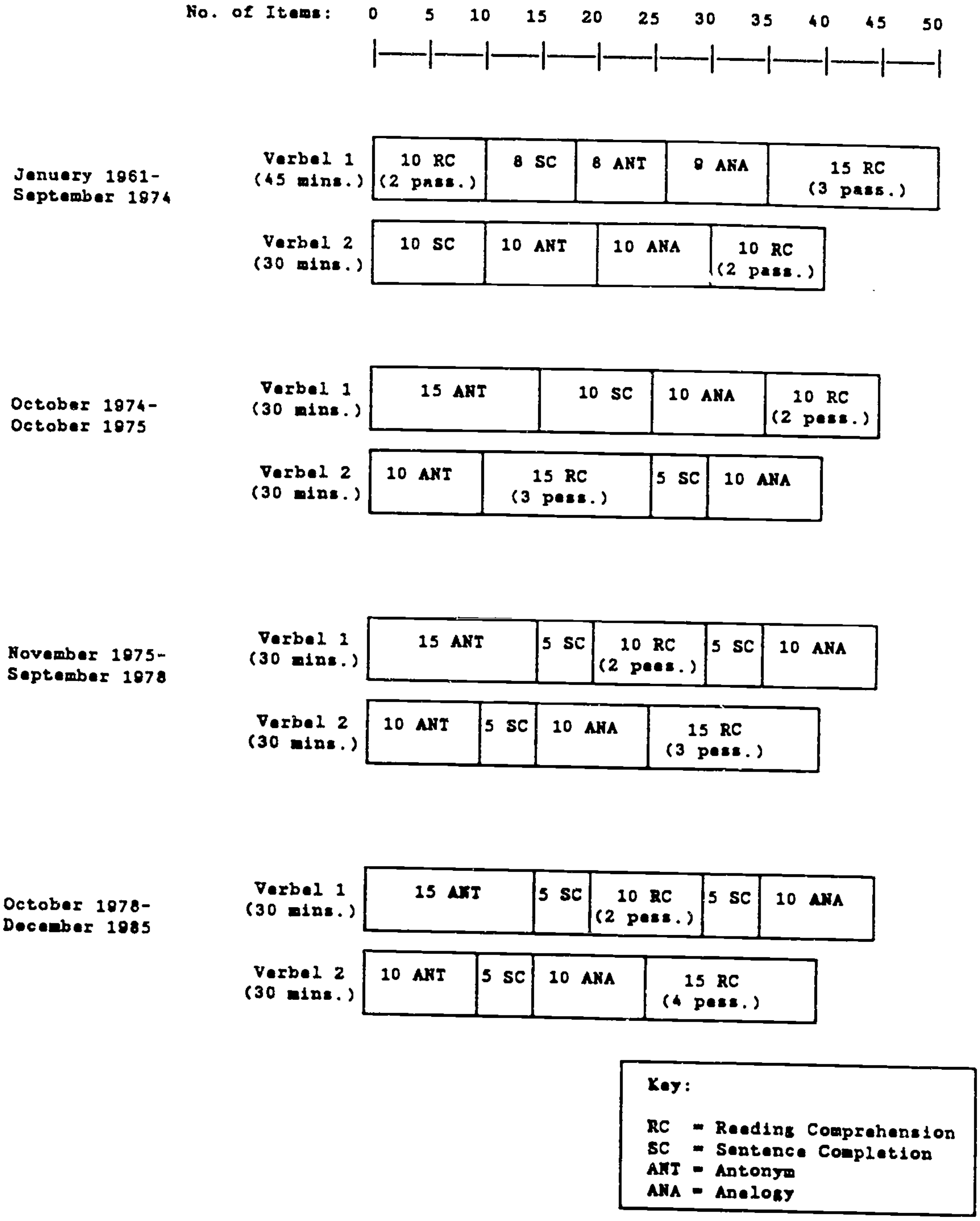
Administration Date	Element	Operational Conversion	Experimental Conversion	Administration Date	Element	Operational Conversion	Experimental Conversion
			1974 Administrations				
November (Change in Specifications)	Type of Equating	Linear	Equipercentile	December (Change in Specifications)	Type of Equating	Linear	Equipercentile
	Raw Score:				Raw Score:		
	70	680	680		70	690	700
	60	610	610		60	620	630
	50	540	540		50	550	550
	40	460	460		40	480	480
	30	390	390		30	410	400
	No. of Cases	398,604	398,604		No. of Cases	231,955	231,955
	Mean	445.86	445.35		Mean	425.11	424.01
	SD	108.83	108.77		SD	107.79	107.24
	Mean Dif. (Exp.-Oper.)	- .51			Mean Dif. (Exp.-Oper.)	- 1.10	
	SD Dif.	6.65			SD Dif.	6.70	
	Root Mean Sq. Dif.	6.67			Root Mean Sq. Dif.	6.79	
	SD Dif. (110)	6.72			SD Dif. (110)	6.83	
	Root Mean Sq. Dif. (110)	6.74			Root Mean Sq. Dif. (110)	6.92	
	Corr. (Oper., Exp.)	.998			Corr. (Oper., Exp.)	.998	
			1982 Administrations				
January Form 1 (No Change in Specifications)	Type of Equating	IRT	Linear	January Form 2 (Change in Specifications)	Type of Equating	IRT	Linear
	Raw Score:				Raw Score:		
	70	660	670		70	650	660
	60	600	600		60	590	600
	50	540	530		50	530	530
	40	470	460		40	470	470
	30	400	400		30	400	400
	No. of Cases	74,130	74,130		No. of Cases	44,844	44,844
	Mean	392.91	394.45		Mean	392.33	394.28
	SD	104.71	102.77		SD	102.68	102.91
	Mean Dif. (Exp.-Oper.)	1.54			Mean Dif. (Exp.-Oper.)	1.95	
	SD Dif.	6.24			SD Dif.	4.49	
	Root Mean Sq. Dif.	6.42			Root Mean Sq. Dif.	4.89	
	SD Dif. (110)	6.55			SD Dif. (110)	4.81	
	Root Mean Sq. Dif. (110)	6.73			Root Mean Sq. Dif. (110)	5.19	
	Corr. (Oper., Exp.)	.998			Corr. (Oper., Exp.)	.999	

Note: Operational and experimental conversions did not take account of "doglegs" used in actual score reporting to ensure that at least one raw score converted to 800. In addition, the standard deviations and root mean squares of differences between conversion lines were calculated in two ways: (1) on actual scores; (2) on scores adjusted to a scale with a standard deviation of 110 for operational scores. The latter values are identified in the table by the "110" in parentheses.

Table 39. Comparisons of Operational and Experimental Conversions for Selected SAT-Mathematical Test Forms

Administration Date	Element	Operational Conversion	Experimental Conversion	Administration Date	Element	Operational Conversion	Experimental Conversion
			1974 Administrations				
November (Change in Specifica-tions)	Type of Equating	Linear	Equipercentile	December (Change in Specifica-tions)	Type of Equating	Linear	Equipercentile
	Raw Score:				Raw Score:		
	50	710	710		50	710	720
	40	620	620		40	620	630
	30	540	540		30	530	530
	20	450	450		20	440	440
	10	360	350		10	350	350
	No. of Cases	398,521	398,521		No. of Cases	231,973	231,973
	Mean	481.08	481.41		Mean	459.96	459.76
	SD	112.20	112.07		SD	112.85	113.36
	Mean Dif. (Exp.-Oper.)	.33			Mean Dif. (Exp.-Oper.)	- .20	
	SD Dif.	5.14			SD Dif.	4.85	
	Root Mean Sq. Dif.	5.15			Root Mean Sq. Dif.	4.86	
	SD Dif. (110)	5.04			SD Dif. (110)	4.73	
	Root Mean Sq. Dif. (110)	5.05			Root Mean Sq. Dif. (110)	4.74	
	Corr. (Oper., Exp.)	.999			Corr. (Oper., Exp.)	.999	
			1982 Administrations				
January Form 1 (No Change in Specifi-cations)	Type of Equating	IRT	Linear	January Form 2 (No Change in Specifi-cations)	Type of Equating	IRT	Linear
	Raw Score:				Raw Score:		
	50	700	680		50	700	700
	40	600	600		40	610	610
	30	510	510		30	520	520
	20	420	430		20	430	440
	10	340	350		10	350	350
	No. of Cases	74,130	74,130		No. of Cases	44,844	44,844
	Mean	435.13	438.11		Mean	432.96	434.99
	SD	115.16	111.40		SD	110.69	109.63
	Mean Dif. (Exp.-Oper.)	2.98			Mean Dif. (Exp.-Oper.)	2.03	
	SD Dif.	7.46			SD Dif.	5.71	
	Root Mean Sq. Dif.	8.03			Root Mean Sq. Dif.	6.06	
	SD Dif. (110)	7.13			SD Dif. (110)	5.67	
	Root Mean Sq. Dif. (110)	7.73			Root Mean Sq. Dif. (110)	6.03	
	Corr. (Oper., Exp.)	.998			Corr. (Oper., Exp.)	.999	

Note: Operational and experimental conversions did not take account of "doglegs" used in actual score reporting to ensure that at least one raw score converted to 800. In addition, the standard deviations and root mean squares of differences between conversion lines were calculated in two ways: (1) on actual scores; (2) on scores adjusted to a scale with a standard deviation of 110 for



No. of Items: 0 5 10 15 20 25 30 35 40 45 50

January 1961-
September 1974

Verbal 1 (45 mins.): 10 RC (2 pass.) | 8 SC | 8 ANT | 9 ANA | 15 RC (3 pass.)

Verbal 2 (30 mins.): 10 SC | 10 ANT | 10 ANA | 10 RC (2 pass.)

October 1974-
October 1975

Verbal 1 (30 mins.): 15 ANT | 10 SC | 10 ANA | 10 RC (2 pass.)

Verbal 2 (30 mins.): 10 ANT | 15 RC (3 pass.) | 5 SC | 10 ANA

November 1975-
September 1978

Verbal 1 (30 mins.): 15 ANT | 5 SC | 10 RC (2 pass.) | 5 SC | 10 ANA

Verbal 2 (30 mins.): 10 ANT | 5 SC | 10 ANA | 15 RC (3 pass.)

October 1978-
December 1985

Verbal 1 (30 mins.): 15 ANT | 5 SC | 10 RC (2 pass.) | 5 SC | 10 ANA

Verbal 2 (30 mins.): 10 ANT | 5 SC | 10 ANA | 15 RC (4 pass.)

Key:

RC = Reading Comprehension
SC = Sentence Completion
ANT = Antonym
ANA = Analogy

Figure 1. Item-Order Specifications for Sections of the SAT-Verbal Test

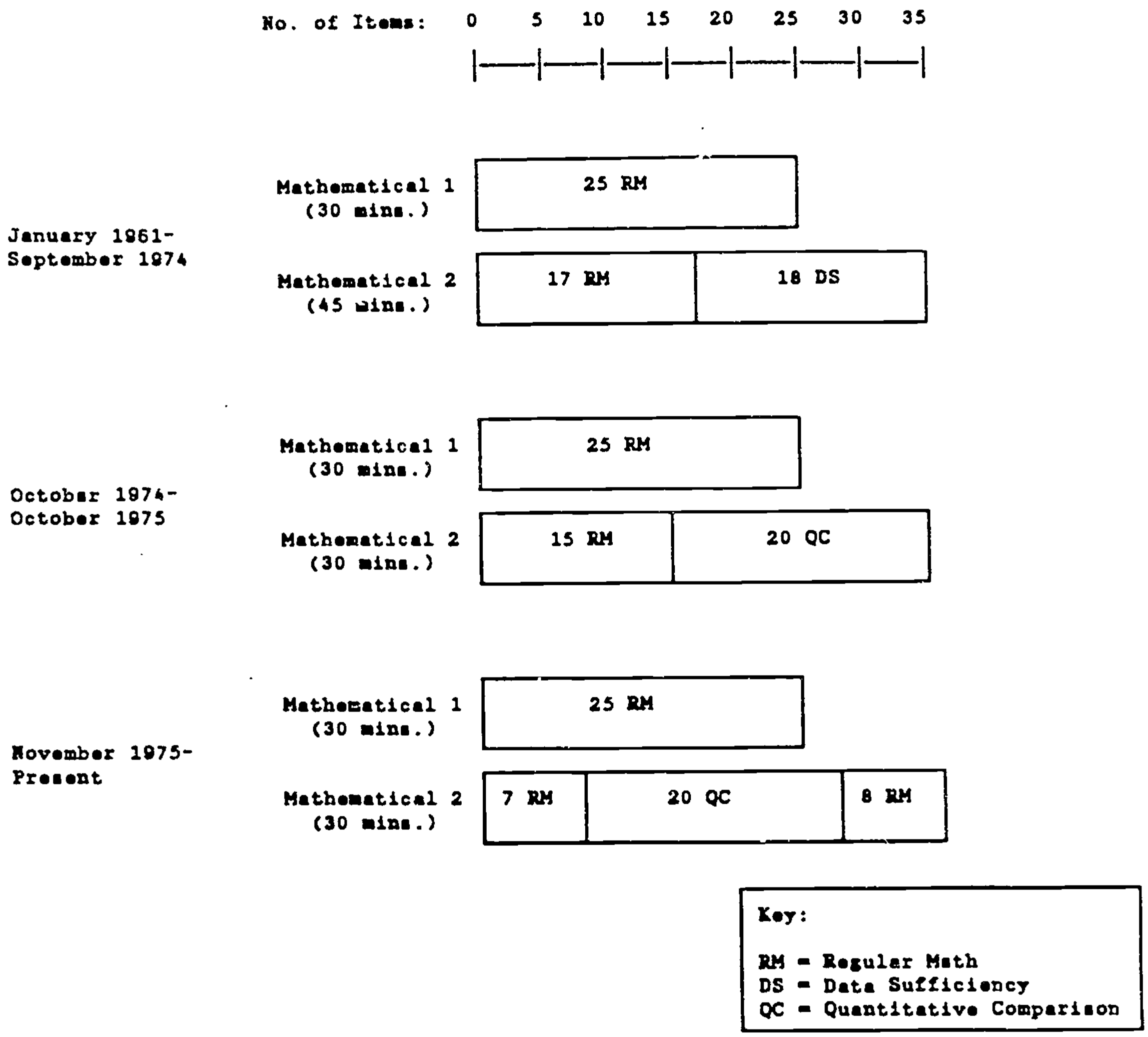
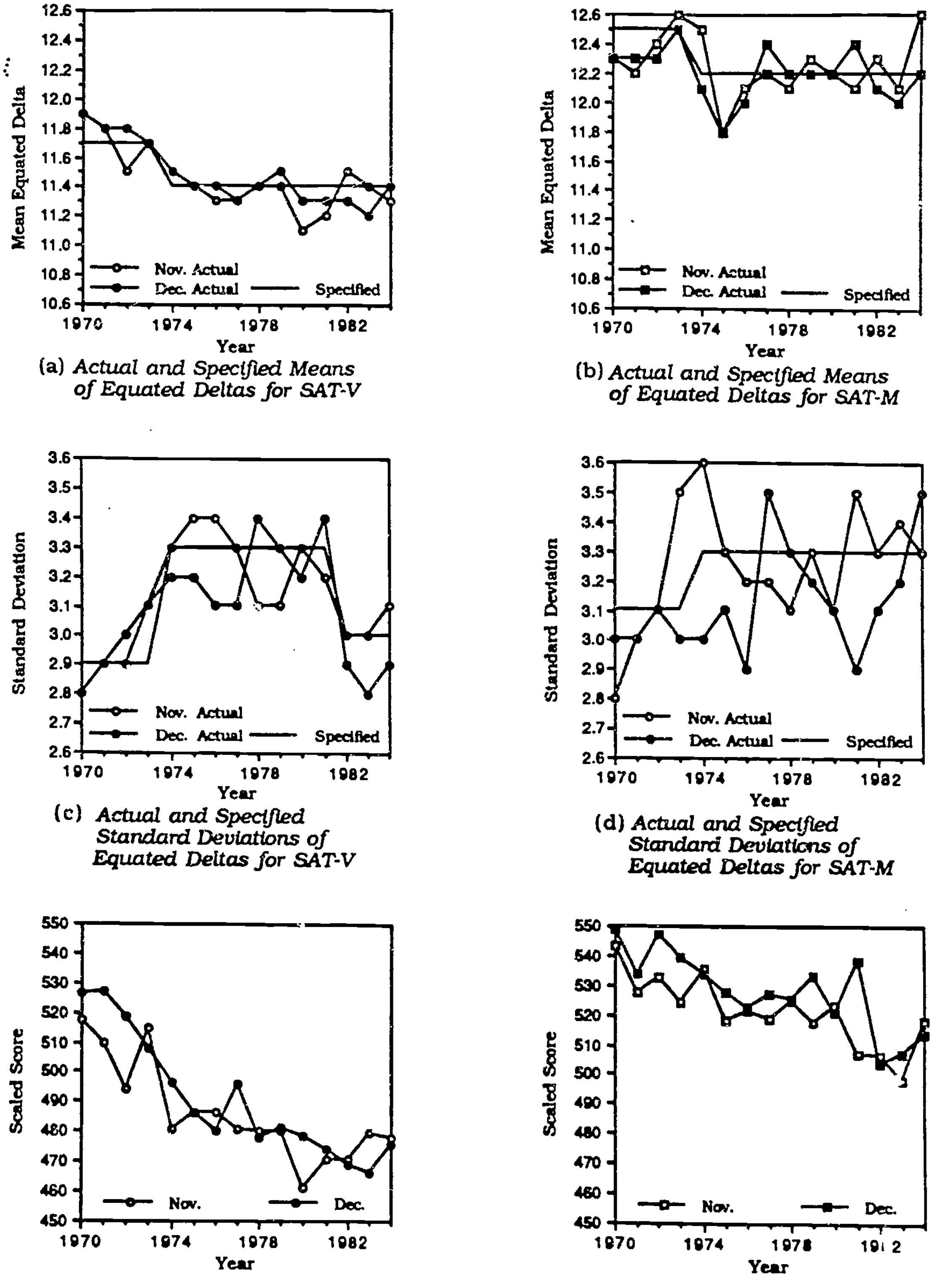


Figure 2. Item-Order Specifications for Sections of the SAT-Mathematical Test



(a) *Actual and Specified Means of Equated Deltas for SAT-V*

(b) *Actual and Specified Means of Equated Deltas for SAT-M*

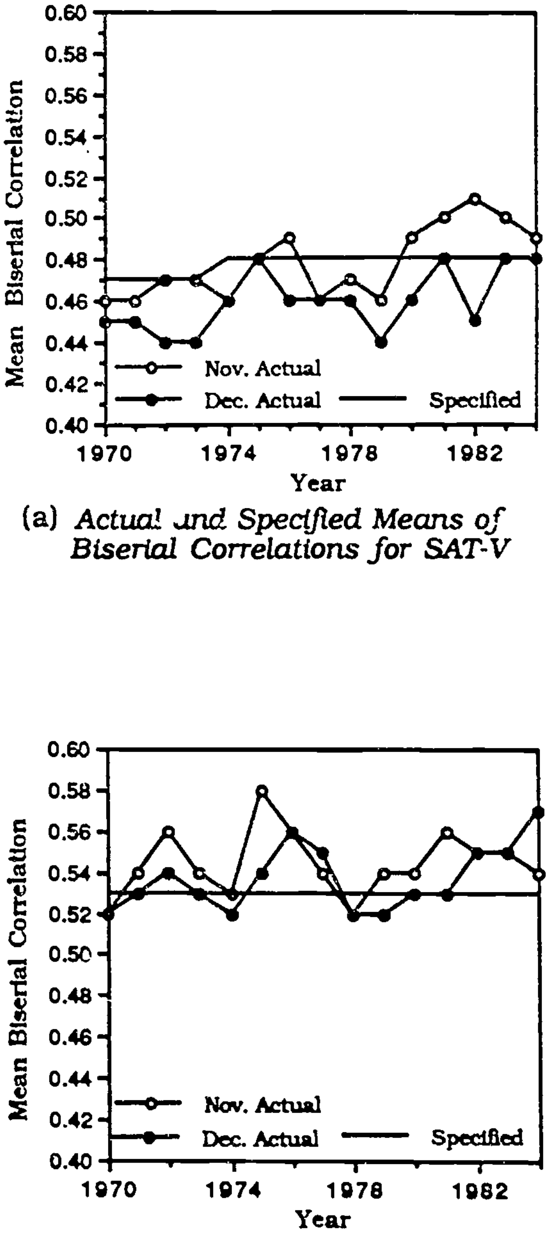
(c) *Actual and Specified Standard Deviations of Equated Deltas for SAT-V*

(d) *Actual and Specified Standard Deviations of Equated Deltas for SAT-M*

(e) *SAT-V Scaled Scores Corresponding to Raw Score Midpoints*

(f) *SAT-M Scaled Scores Corresponding to Raw Score Midpoints*

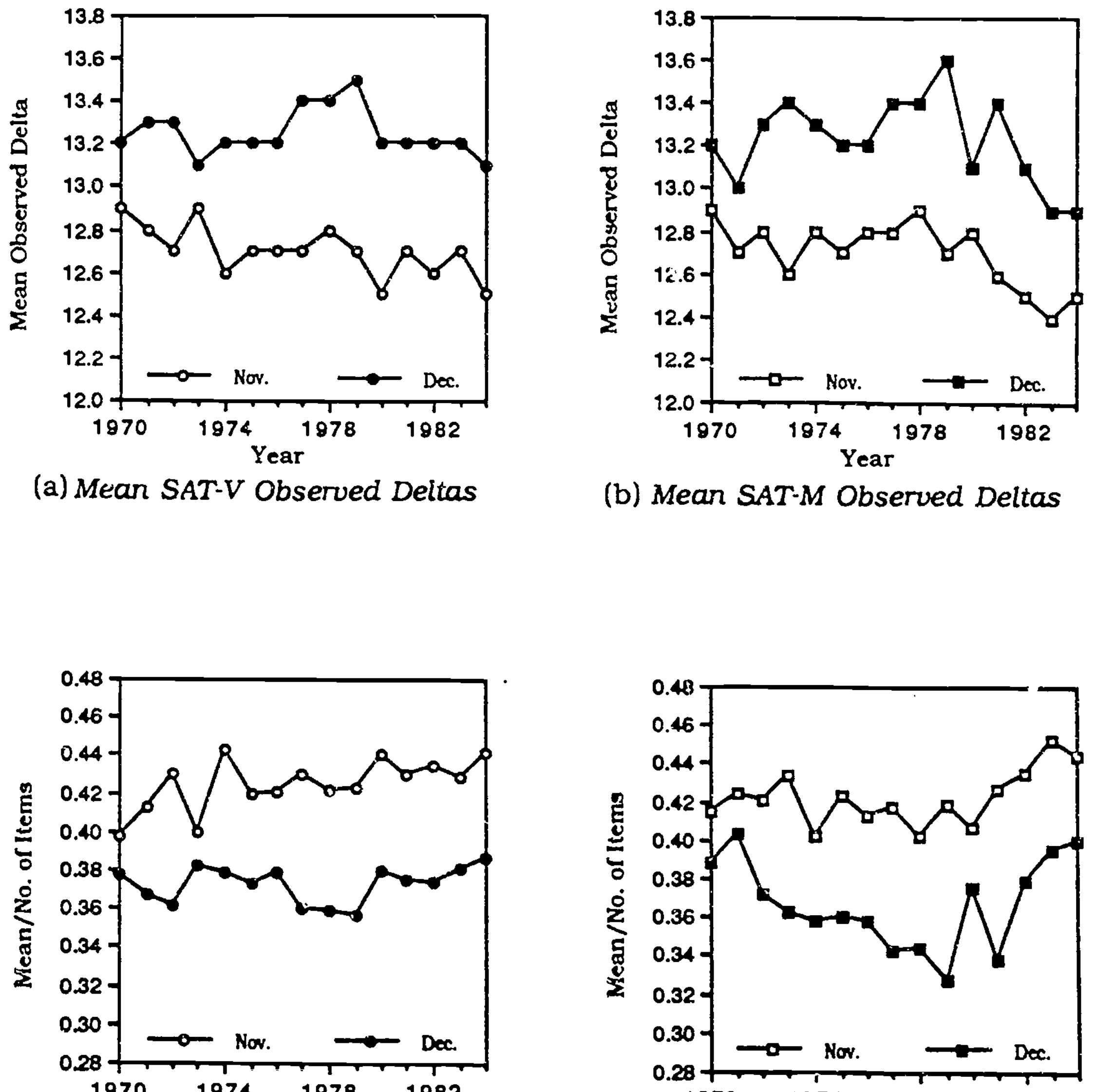
Figure 3. Trends in Test Difficulty for November and December SAT Test Forms Administered from 1970 to 1984



(a) *Actual and Specified Means of Biserial Correlations for SAT-V*

(b) *Actual and Specified Means of Biserial Correlations for SAT-M*

Figure 4. Trends in Actual and Specified Biserial Correlations for November and December SAT Test Forms from 1970 to 1984



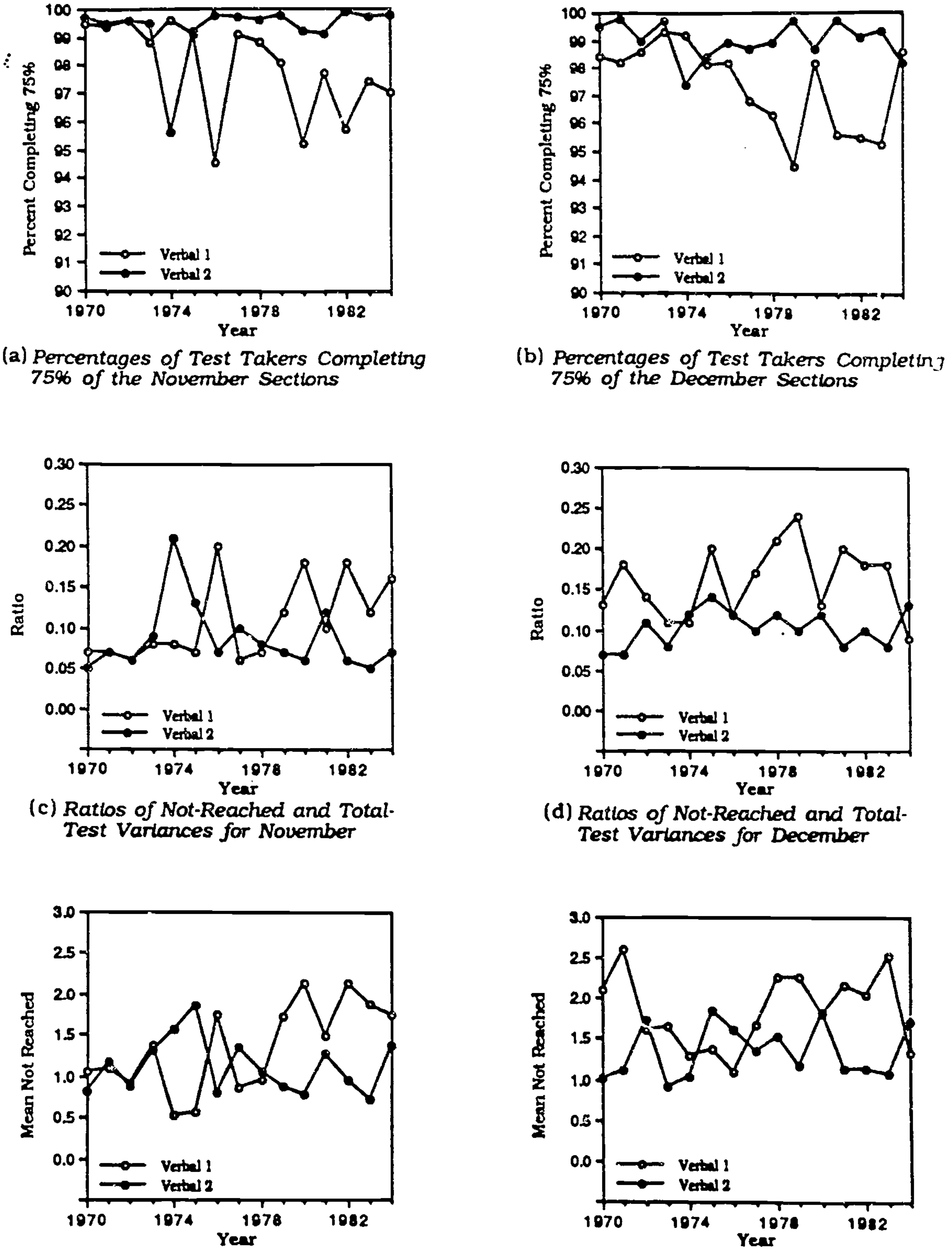
(a) *Mean SAT-V Observed Deltas*

(b) *Mean SAT-M Observed Deltas*

(c) *SAT-V Raw-Score Means Divided by Number of Items*

(d) *SAT-M Raw-Score Means Divided by Number of Items*

Figure 5. Trends in Relative Test Difficulty for November and December SAT Test Forms Administered from 1970 to 1984



(a) *Percentages of Test Takers Completing 75% of the November Sections*

(b) *Percentages of Test Takers Completing 75% of the December Sections*

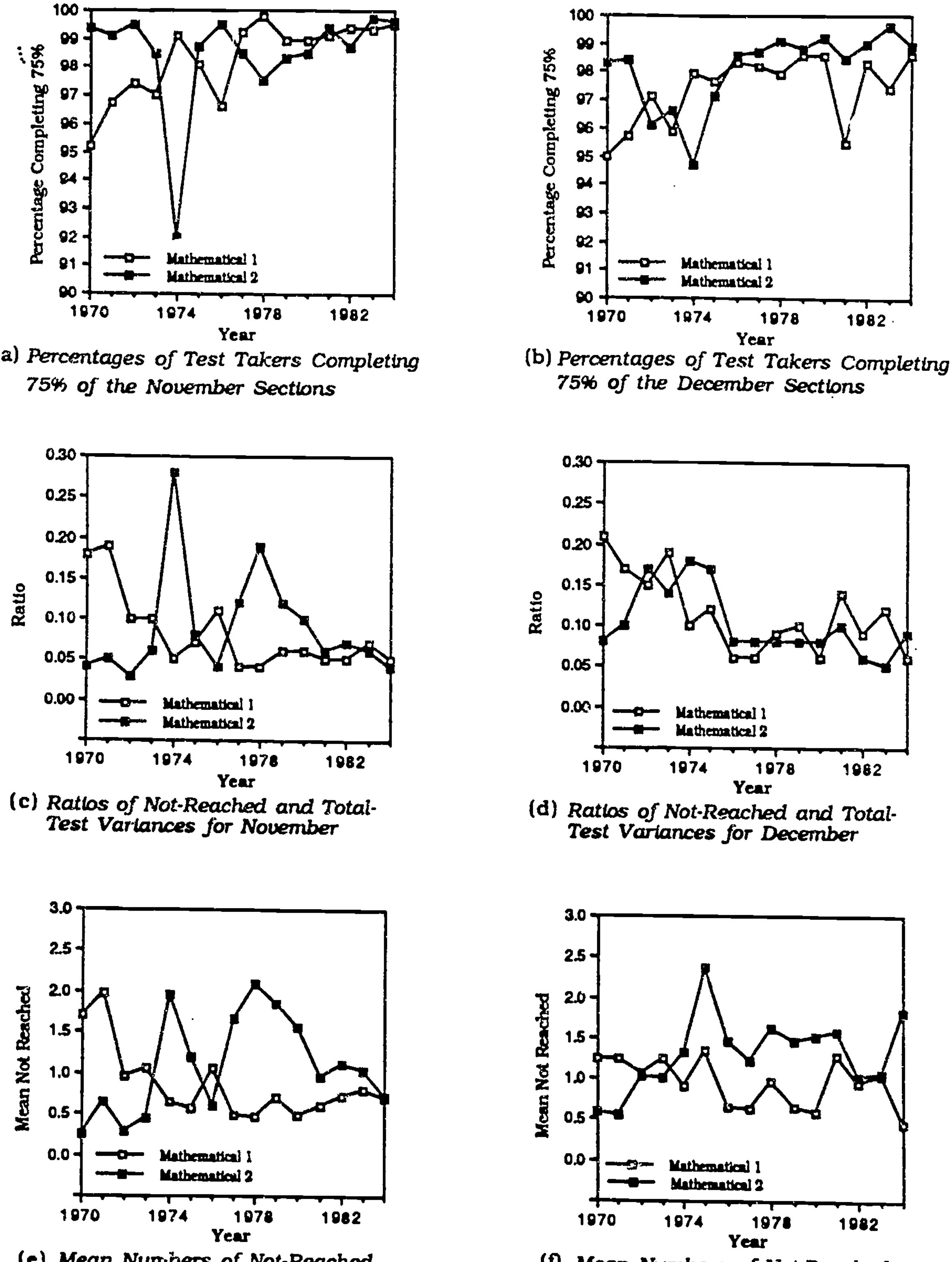
(c) *Ratios of Not-Reached and Total-Test Variances for November*

(d) *Ratios of Not-Reached and Total-Test Variances for December*

(e) *Mean Numbers of Not-Reached Items for November*

(f) *Mean Numbers of Not-Reached Items for December*

Figure 6. Trends in Speededness Indices for November and December SAT-Verbal Test Forms Administered From 1970 to 1984



(a) *Percentages of Test Takers Completing 75% of the November Sections*

(b) *Percentages of Test Takers Completing 75% of the December Sections*

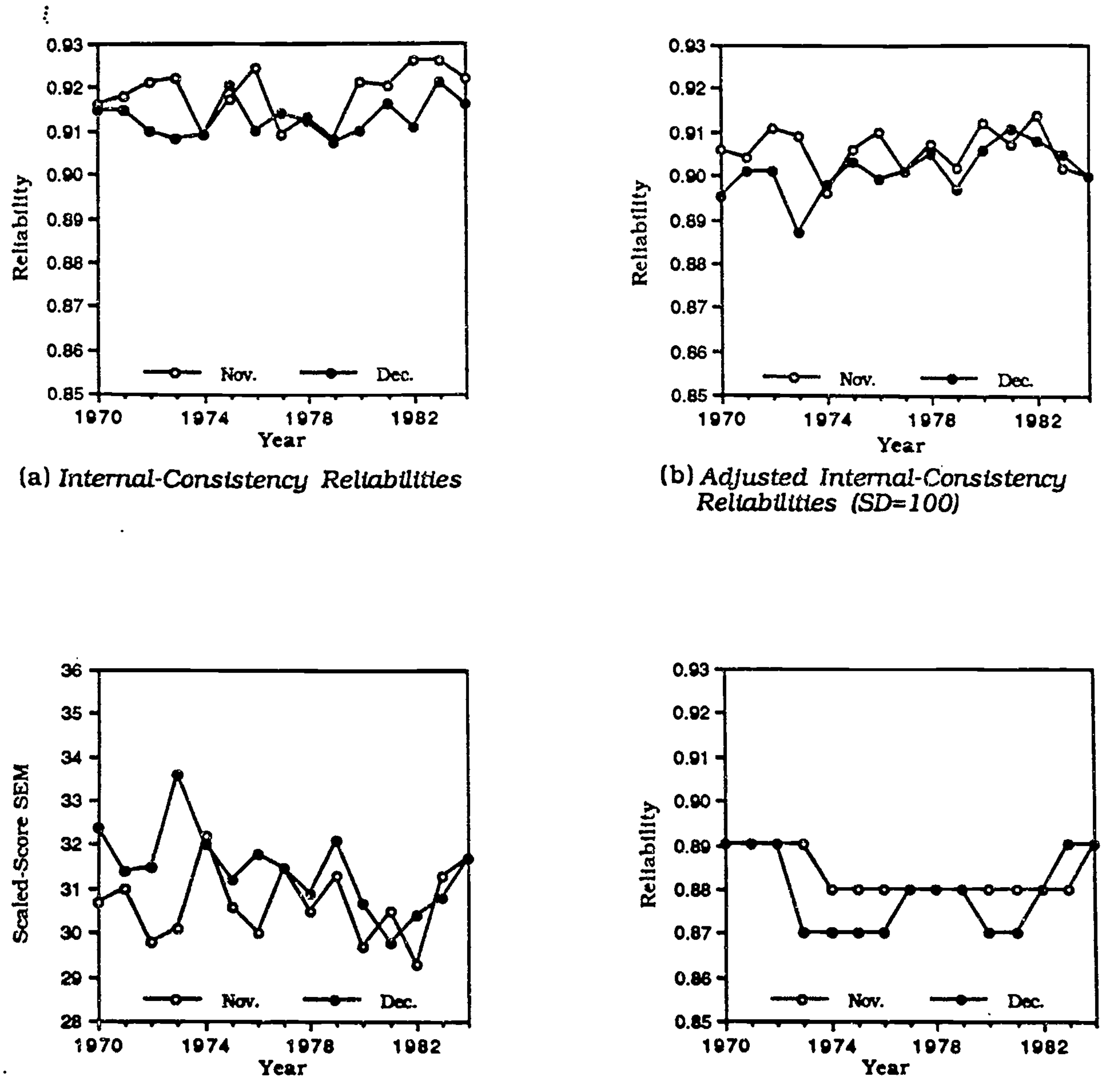
(c) *Ratios of Not-Reached and Total-Test Variances for November*

(d) *Ratios of Not-Reached and Total-Test Variances for December*

(e) *Mean Numbers of Not-Reached Items for November*

(f) *Mean Numbers of Not-Reached Items for December*

Figure 7. Trends in Speededness Indices for November and December SAT-Mathematical Test Forms Administered from 1970 to 1984



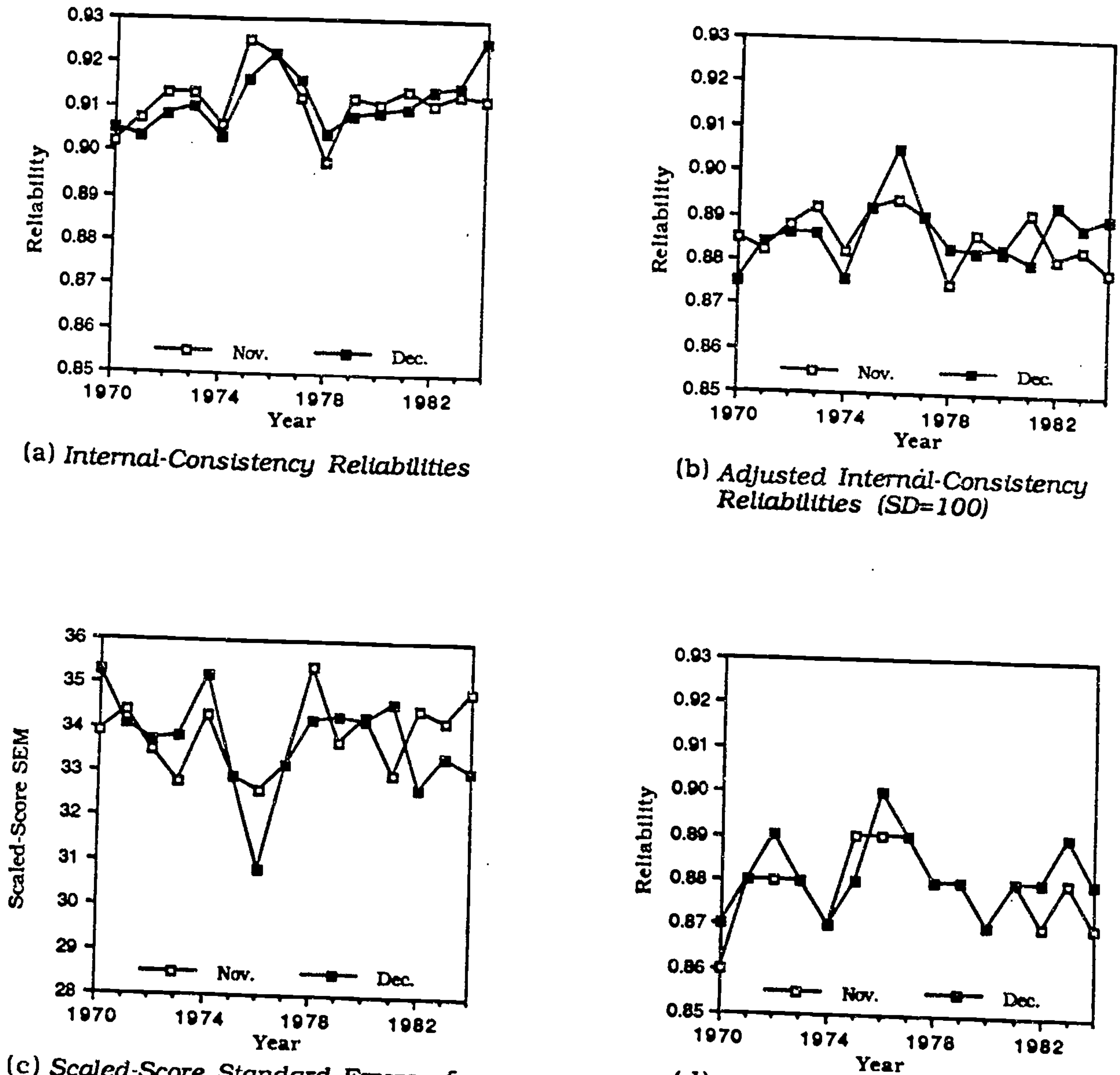
(a) *Internal-Consistency Reliabilities*

(b) *Adjusted Internal-Consistency Reliabilities (SD=100)*

(c) *Scaled-Score Standard Errors of Measurement*

(d) *Test-Retest Reliabilities (for Large-Volume Repeater Patterns)*

Figure 8. Trends in Reliability Information for November and December SAT-Verbal Test Forms Administered from 1970 to 1984



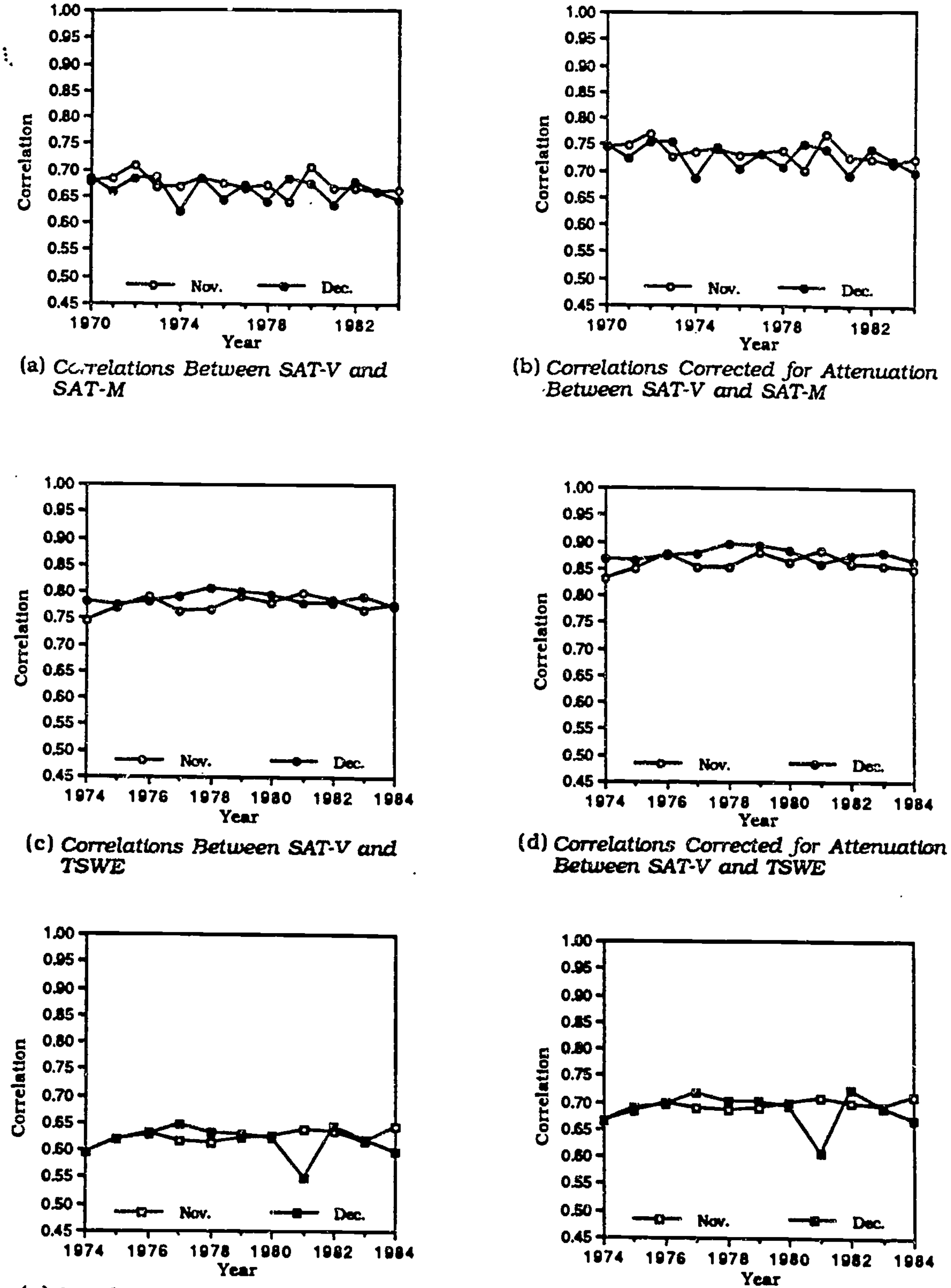
(a) *Internal-Consistency Reliabilities*

(b) *Adjusted Internal-Consistency Reliabilities (SD=100)*

(c) *Scaled-Score Standard Errors of Measurement*

(d) *Test-Retest Reliabilities (for Large-Volume Repeater Patterns)*

Figure 9. Trends in Reliability Information for November and December SAT-Mathematical Test Forms Administered from 1970 to 1984



(a) Cc.rrelations Between SAT-V and SAT-M

(b) Correlations Corrected for Attenuation Between SAT-V and SAT-M

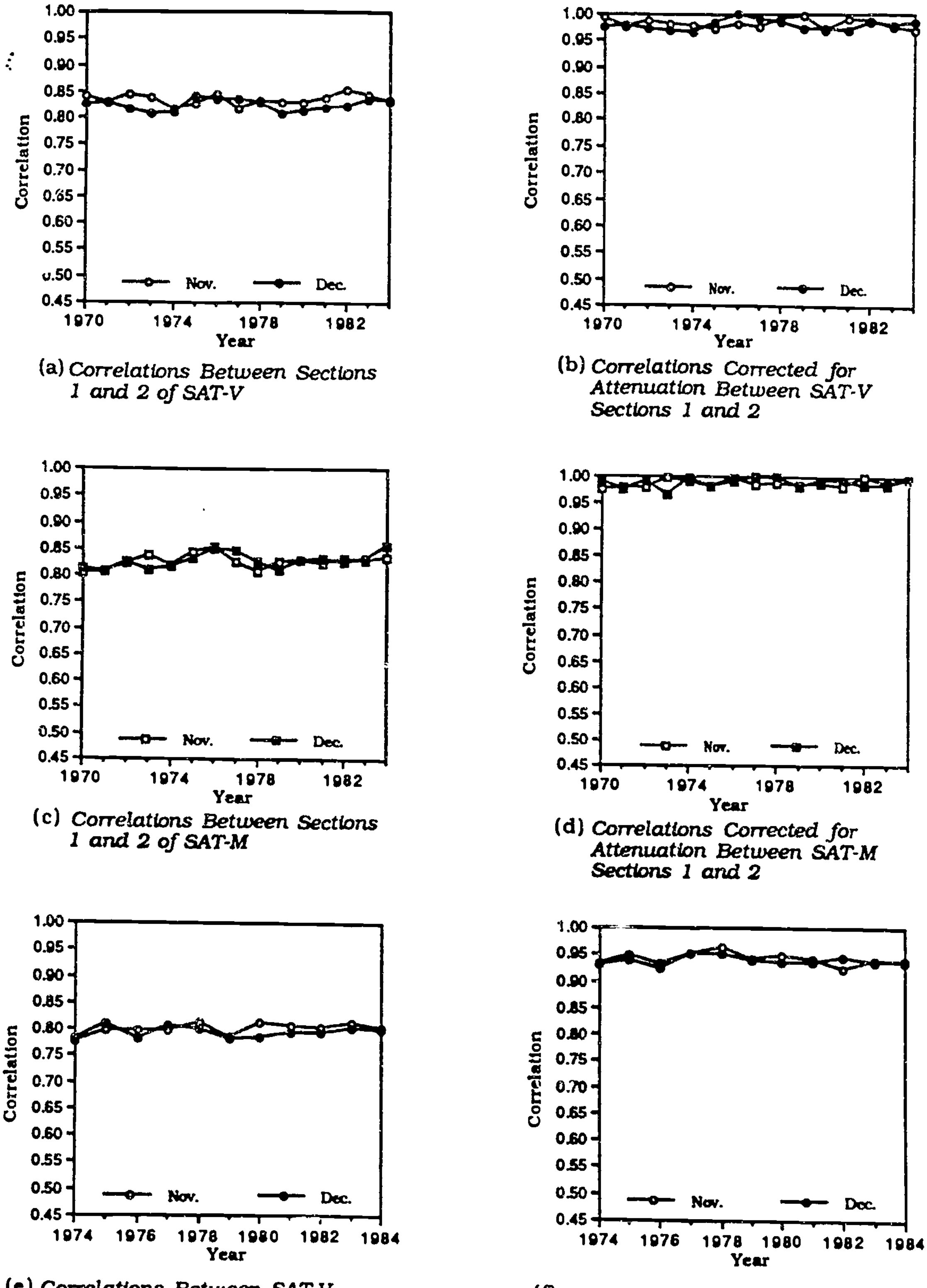
(c) Correlations Between SAT-V and TSWE

(d) Correlations Corrected for Attenuation Between SAT-V and TSWE

(e) Correlations Between SAT-M and TSWE

(f) Correlations Corrected for Attenuation Between SA'-M and TSWE

Figure 10. Trends in Correlational Patterns Among SAT-Verbal, SAT-Mathematical, and TSWE for November and December Test Forms Adminstered from 1970 to 1984



(a) *Correlations Between Sections 1 and 2 of SAT-V*

(b) *Correlations Corrected for Attenuation Between SAT-V Sections 1 and 2*

(c) *Correlations Between Sections 1 and 2 of SAT-M*

(d) *Correlations Corrected for Attenuation Between SAT-M Sections 1 and 2*

(e) *Correlations Between SAT-V Reading and Vocabulary Subscores*

(f) *Correlations Corrected for Attenuation Between SAT Reading and Vocabulary Subscores*

Figure 11. Trends in Correlational Patterns for Verbal Sections, Mathematical Sections, and Verbal Subscores for November and December SAT Test Forms Administered from 1970 to 1984

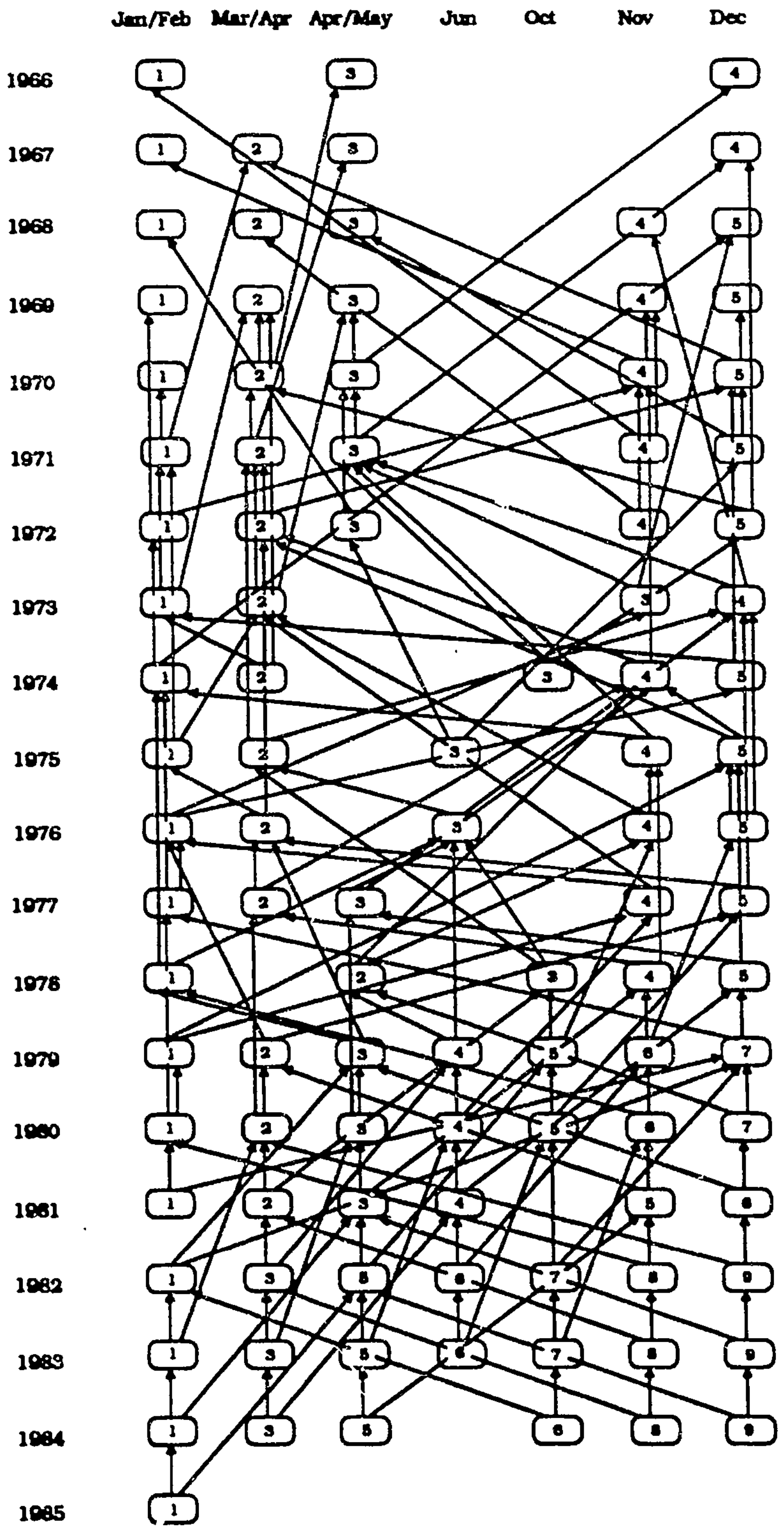
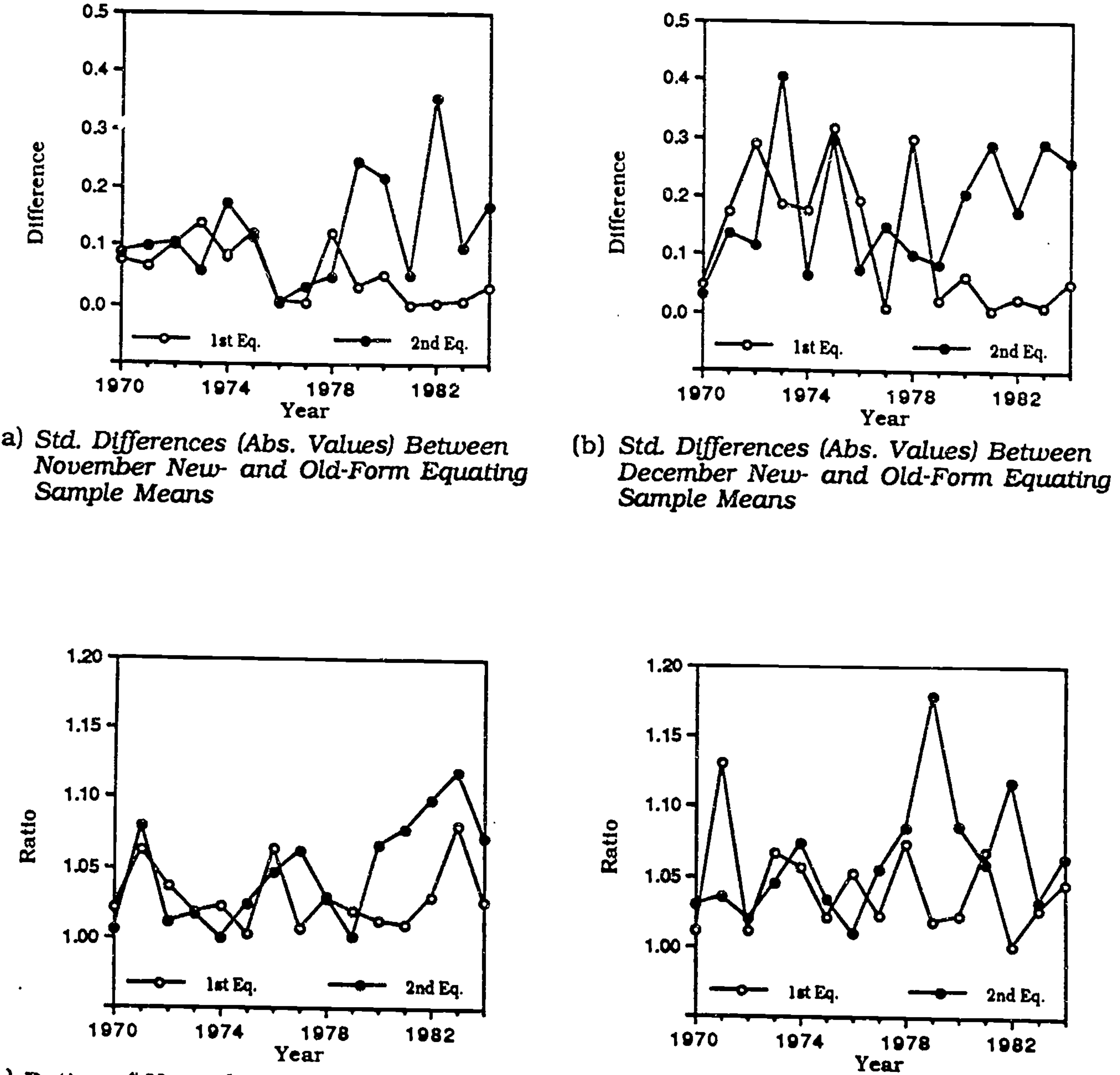


Figure 12. Genealogical Chart for SAT-Verbal (This figure does not include new or revised forms that lie outside the braiding plan.)



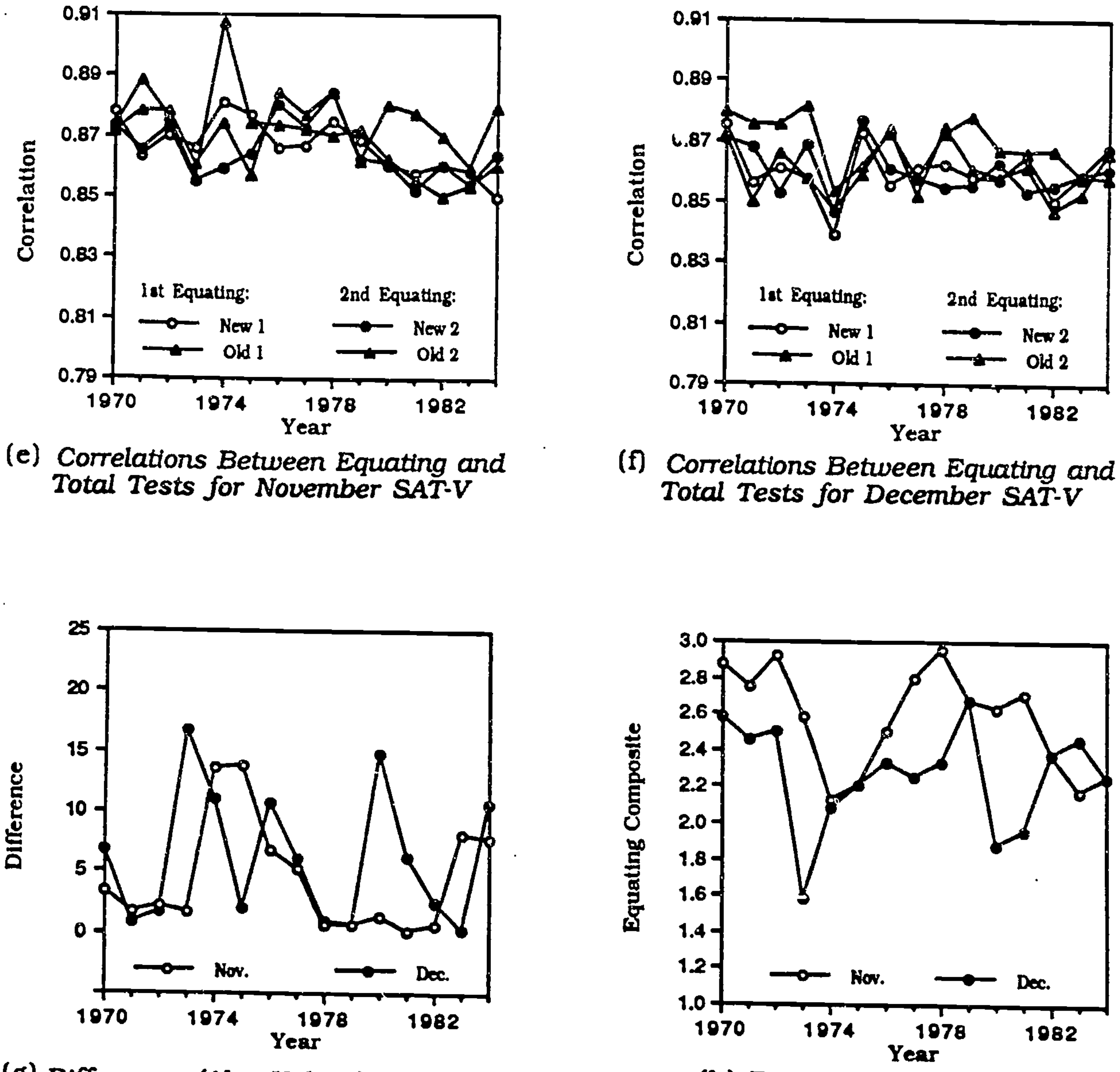
(a) *Std. Differences (Abs. Values) Between November New- and Old-Form Equating Sample Means*

(b) *Std. Differences (Abs. Values) Between December New- and Old-Form Equating Sample Means*

(c) *Ratios of November New- and Old-Form Equating Sample Variances*

(d) *Ratios of December New- and Old-Form Equating Sample Variances*

Figure 13. Trends in Equating Information for November and December SAT-Verbal Test Forms Administered from 1970 to 1984



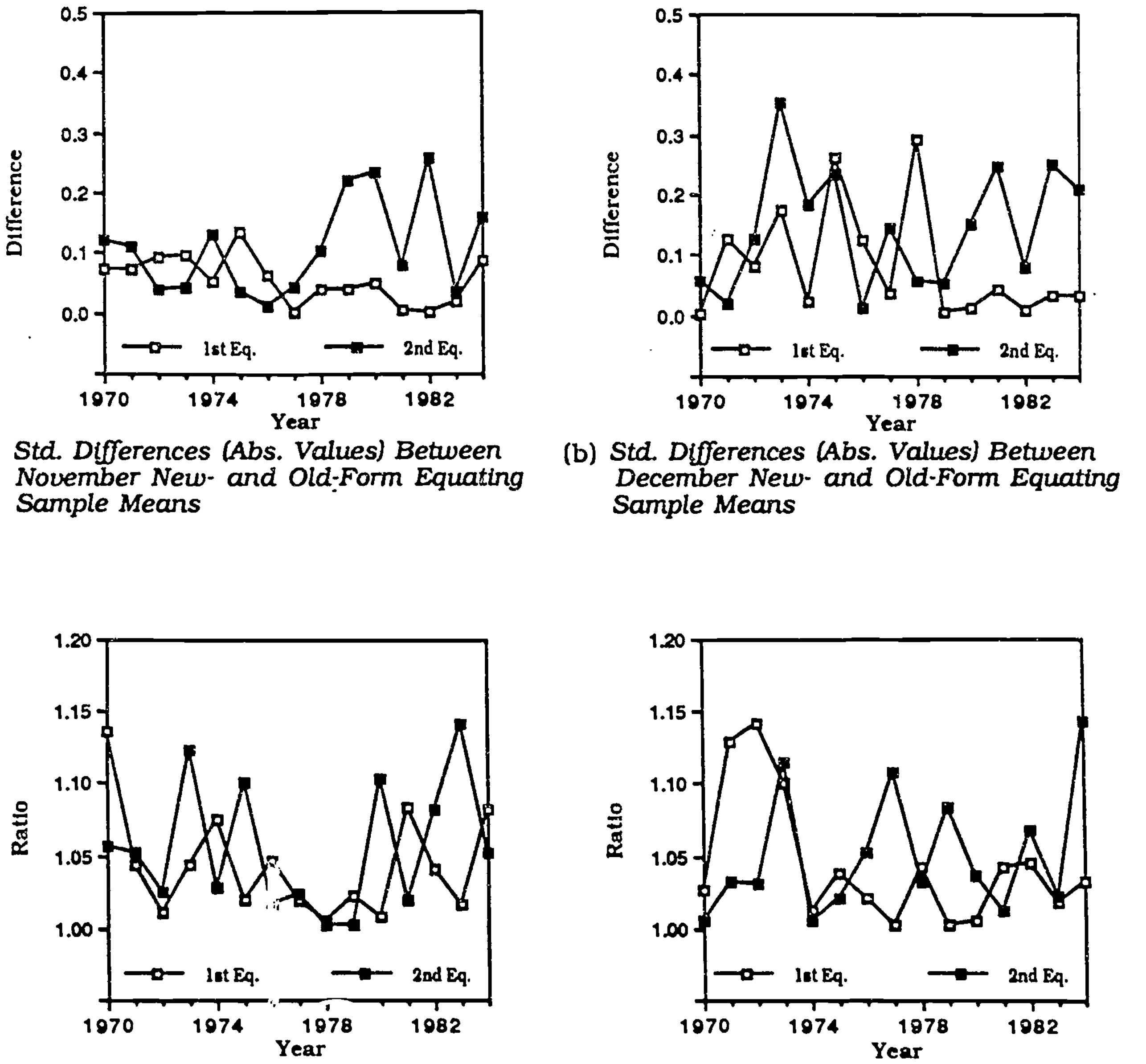
(e) *Correlations Between Equating and Total Tests for November SAT-V*

(f) *Correlations Between Equating and Total Tests for December SAT-V*

(g) *Differences (Abs. Values) Between 1st and 2nd Equating Conversion Lines at Midpoint*

(h) *Equating Composites*

Figure 13. (continued)



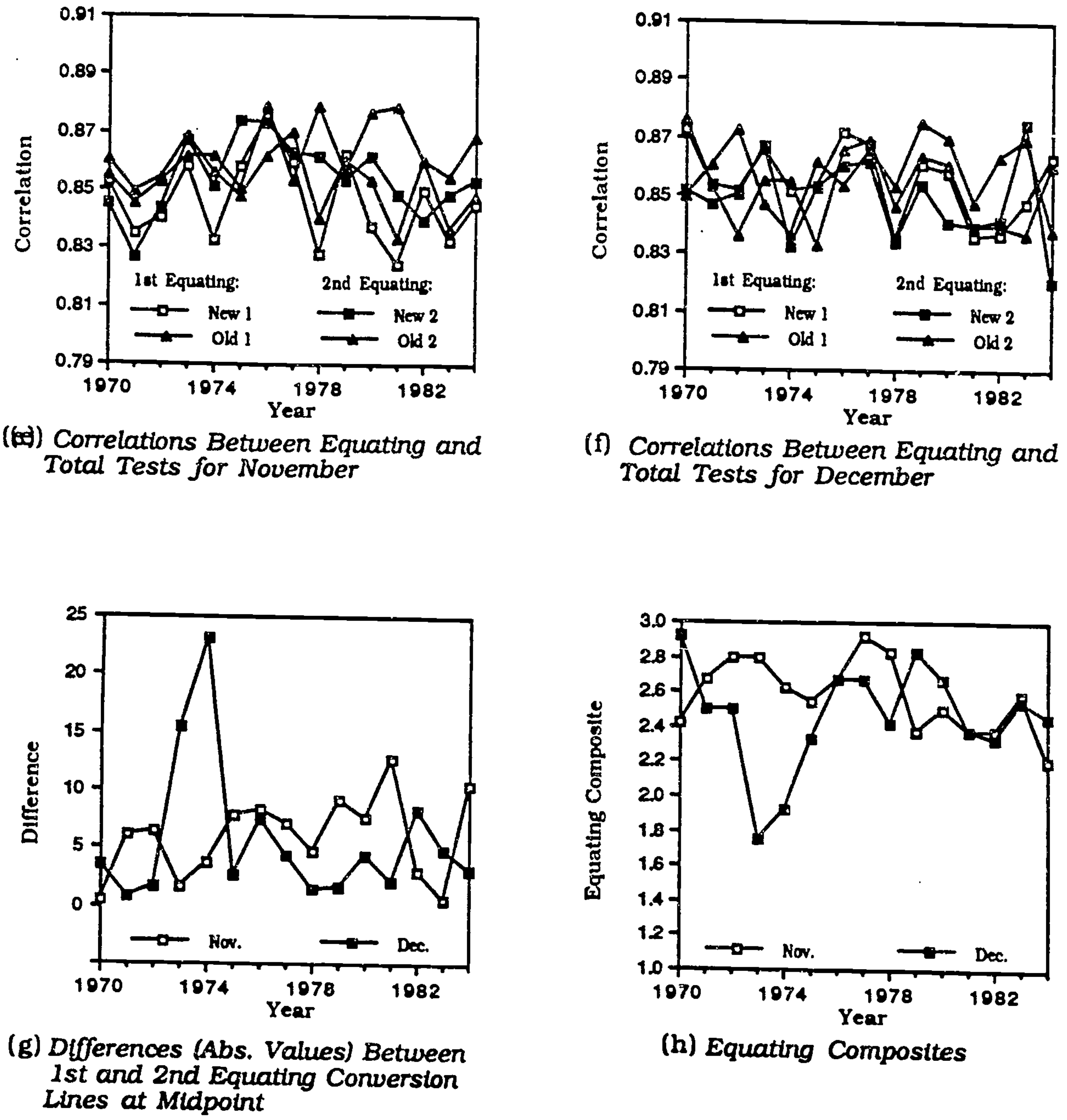
(a) *Std. Differences (Abs. Values) Between November New- and Old-Form Equating Sample Means*

(b) *Std. Differences (Abs. Values) Between December New- and Old-Form Equating Sample Means*

(c) *Ratios of November New- and Old-Form Equating Sample Variances*

(d) *Ratios of December New- and Old-Form Equating Sample Variances*

Figure 14. Trends in Equating Information for November and December SAT-Mathematical Test Forms Administered from 1970 tr 1984



(e) *Correlations Between Equating and Total Tests for November*

(f) *Correlations Between Equating and Total Tests for December*

(g) *Differences (Abs. Values) Between 1st and 2nd Equating Conversion Lines at Midpoint*

(h) *Equating Composites*

Figure 14. (continued)

APPENDIX A

SAT DIRECTIONS AND SAMPLE QUESTIONS

SAT-VERBAL

Antonyms, Analogies, Sentence Completions, Reading Comprehension

SECTION **1**	Time—30 minutes 40 Questions	For each question in this section, choose the best answer and fill in the corresponding oval on the answer sheet.

Antonyms

Each question below consists of a word in capital letters, followed by five lettered words or phrases. Choose the word or phrase that is most nearly opposite in meaning to the word in capital letters. Since some of the questions require you to distinguish fine shades of meaning, consider all the choices before deciding which is best.

Example:

GOOD: (A) sour (B) bad (C) red
(D) hot (E) ugly

Ⓐ ● Ⓒ Ⓓ Ⓔ

Sample Questions

1. SURPLUS : (A) shortage (B) criticism
(C) heated argument (D) sudden victory
(E) thorough review

2. TEMPESTUOUS : (A) responsible
(B) predictable (C) tranquil
(D) prodigious (E) tentative

Correct Answers: 1. A
2. C

Analogies

Each question below consists of a related pair of words or phrases, followed by five lettered pairs of words or phrases. Select the lettered pair that best expresses a relationship similar to that expressed in the original pair.

Example:

YAWN : BOREDOM :: (A) dream : sleep
(B) anger : madness (C) smile : amusement
(D) face : expression (E) impatience : rebellion

Ⓐ Ⓑ ● Ⓓ Ⓔ

Sample Questions

3. APPAREL : SHIRT :: (A) sheep : wool
(B) foot : shoe (C) light : camera
(D) belt : buckle (E) jewelry : ring

4. BUNGLER : SKILL :: (A) fool : amusement
(B) critic : error (C) daredevil : caution
(D) braggart : confidence (E) genius : intelligence

Correct Answers: 3. E
4. C

Sentence Completions

Each sentence below has one or two blanks, each blank indicating that something has been omitted. Beneath the sentence are five lettered words or sets of words. Choose the word or set of words that, when inserted in the sentence, best fits the meaning of the sentence as a whole.

Example:

Although its publicity has been ----, the film itself is intelligent, well-acted, handsomely produced, and altogether ----.

(A) tasteless. .respectable (B) extensive. .moderate (C) sophisticated. .amateur (D) risqué. .crude (E) perfect. .spectacular

● Ⓑ Ⓒ Ⓓ Ⓔ

Sample Questions

5. Either the sunsets at Nome are ----, or the one I saw was a poor example.

   (A) gorgeous (B) overrated (C) unobserved (D) exemplary (E) unappreciated

6. Specialization has been emphasized to such a degree that some students ---- nothing that is ---- to their primary area of interest.

   (A) ignore. .contradictory
   (B) incorporate. .necessary
   (C) recognize. .fundamental
   (D) accept. .relevant
   (E) value. .extraneous

Correct Answers: 5. B
6. E

Reading Comprehension

Each passage below is followed by questions based on its content. Answer the questions following each passage on the basis of what is stated or implied in that passage.

From the beginning, this trip to the high plateaus in Utah has had the feel of a last visit. We are getting beyond the age when we can unroll our sleeping bags under any pine or in any wash, and the gasoline situation throws the future of automobile touring into doubt. I would hate to have missed the extravagant personal liberty that wheels and cheap gasoline gave us, but I will not mourn its passing. It was part of our time of wastefulness and excess. Increasingly, we will have to earn our admission to this spectacular country. We will have to come by bus, as foreign tourists do, and at the end of the bus line use our legs. And if that reduces the number of people who benefit every year, the benefit will be qualitatively greater, for what most recommends the plateaus and their intervening deserts is not people, but space, emptiness, silence, awe.

I could make a suggestion to the road builders, too. The experience of driving the Aquarius Plateau on pavement is nothing like so satisfying as the old experience of driving it on rocky, rutted, chuckholed, ten-mile-an-hour dirt. The road will be a lesser thing when it is paved all the way, and so will the road over the Fish Lake Hightop, and the one over the Wasatch Plateau, and the steep road over the Tushar, the highest of the plateaus, which we will travel tomorrow. To substitute comfort and ease for real experience is too American a habit to last. It is when we feel the earth rough to all our length, as in Robert Frost's poem, that we know it as its creatures ought to know it.

The reading passages in this test are brief excerpts or adaptations of excerpts from published material. The ideas contained in them do not necessarily represent the opinions of the College Board or Educational Testing Service. To make the text suitable for testing purposes, we may in some cases have altered the style, contents, or point of view of the original.

7. According to the author, what will happen if fewer people visit the high country each year?

   (A) The characteristic mood of the plateaus will be tragically altered.
   (B) The doctrine of personal liberty will be seriously undermined.
   (C) The pleasure of those who do go will be heightened.
   (D) The people who visit the plateaus will have to spend more for the trip.
   (E) The paving of the roads will be slowed down considerably.

8. The author most probably paraphrases part of a Robert Frost poem in order to

   (A) lament past mistakes
   (B) warn future generations
   (C) reinforce his own sentiments
   (D) show how poetry enhances civilization
   (E) emphasize the complexity of the theme

9. It can be inferred from the passage that the author regards the paving of the plateau roads as

   (A) a project that will never be completed
   (B) a conscious attempt to destroy scenic beauty
   (C) an illegal action
   (D) an inexplicable decision
   (E) an unfortunate change

Correct Answers: 7. C
8. C
9. E

SAT- MATHEMATICAL

Regular Mathematics, Data Sufficiency, Quantitative Comparisons

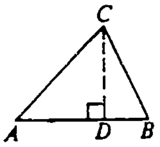
SECTION **2** Time—30 minutes 25 Questions

In this section solve each problem, using any available space on the page for scratchwork. Then decide which is the best of the choices given and fill in the corresponding oval on the answer sheet.

The following information is for your reference in solving some of the problems.

Circle of radius r: Area $= \pi r^2$; Circumference $= 2\pi r$
The number of degrees of arc in a circle is 360.
The measure in degrees of a straight angle is 180.

Definition of symbols:
$=$ is equal to
$\neq$ is unequal to
$<$ is less than
$>$ is greater than
$\leq$ is less than or equal to
$\geq$ is greater than or equal to
$\parallel$ is parallel to
$\perp$ is perpendicular to



Triangle: The sum of the measures in degrees of the angles of a triangle is 180.

If $\angle CDA$ is a right angle, then

(1) area of $\triangle ABC = \frac{AB \times CD}{2}$

(2) $AC^2 = AD^2 + DC^2$

Note: Figures that accompany problems in this test are intended to provide information useful in solving the problems. They are drawn as accurately as possible EXCEPT when it is stated in a specific problem that its figure is not drawn to scale. All figures lie in a plane unless otherwise indicated. All numbers used are real numbers.

Regular Mathematics

Sample Questions

1. If $2y = 3$, then $3(2y)^2 =$

(A) $\frac{27}{4}$

(B) 18

(C) $\frac{81}{4}$

(D) 27

(E) 81

2. Of seven consecutive integers in increasing order, if the sum of the first three integers is 33, what is the sum of the last three integers?

(A) 36
(B) 39
(C) 42
(D) 45
(E) 48

Correct Answers: 1. D
2. D

Data Sufficiency

Directions: Each of the data sufficiency problems below consists of a question and two statements, labeled (1) and (2), in which certain data are given. You have to decide whether the data given in the statements are sufficient for answering the question. Using the data given in the statements plus your knowledge of mathematics and everyday facts (such as the number of days in July or the meaning of counterclockwise), you are to fill in the corresponding oval

A if statement (1) ALONE is sufficient, but statement (2) alone is not sufficient to answer the question asked;
B if statement (2) ALONE is sufficient, but statement (1) alone is not sufficient to answer the question asked;
C if BOTH statements (1) and (2) TOGETHER are sufficient to answer the question asked, but NEITHER statement ALONE is sufficient;
D if EACH statement ALONE is sufficient to answer the question asked;
E if statements (1) and (2) TOGETHER are NOT sufficient to answer the question asked, and additional data specific to the problem are needed.

Numbers: All numbers used are real numbers.

Figures: A figure in a data sufficiency problem will conform to the information given in the question, but will not necessarily conform to the additional information given in statements (1) and (2).

You may assume that lines shown as straight are straight and that angle measures are greater than zero.

You may assume that the position of points, angles, regions, etc., exist in the order shown.

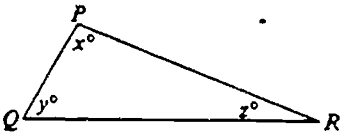
All figures lie in a plane unless otherwise indicated.

Example:

In $\triangle PQR$, what is the value of x?

(1) $PQ = PR$

(2) $y = 40$



Explanation: According to statement (1), $PQ = PR$; therefore, $\triangle PQR$ is isosceles and $y = z$. Since $x + y + z = 180$, $x + 2y = 180$. Since statement (1) does not give a value for y, you cannot answer the question using statement (1) by itself. According to statement (2), $y = 40$; therefore, $x + z = 140$. Since statement (2) does not give a value for z, you cannot answer the question using statement (2) by itself. Using both statements together, you can find y and z; therefore, you can find x, and the answer to the problem is C.

Sample Questions

3. Is $a + b = a$?

(1) $b = 0$

(2) $a = 10$

4. Is rectangle R a square?

(1) The area of R is 16.

(2) The length of a side of R is 4.

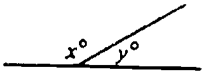
Correct Answers: 3. A
4. C

Questions 5-6 each consist of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet fill in oval

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined from the information given.

AN E RESPONSE WILL NOT BE SCORED.

EXAMPLES

	Column A	Column B	Answers
E1.	2×6	$2 + 6$	● ○ ○ ○ ○
			
E2.	$180 - x$	y	○ ○ ● ○ ○
E3.	$p - q$	$q - p$	○ ○ ○ ● ○

Notes:

1. In certain questions, information concerning one or both of the quantities to be compared is centered above the two columns.
2. In a given question, a symbol that appears in both columns represents the same thing in Column A as it does in Column B.
3. Letters such as x, n, and k stand for real numbers.

Sample Questions

	Column A	Column B
5.	The least positive integer divisible by 2, 3, and 4	24

Parallel lines ℓ_1 and ℓ_2 are 2 inches apart. P is a point on ℓ_1 and Q is a point on ℓ_2.

6.	Length of PQ	3 inches

Correct Answers: 5. B
6. D

EXAMPLES OF EXPLAINED SAT ITEMS•

Analogy Example

Remember that a pair of words can have more than one relationship. For example:

PRIDE : LION : : (A) snake : python (B) pack : wolf
(C) rat : mouse (D) bird : starling (E) dog : canine

A possible relationship between *pride* and *lion* might be that "the first word describes a characteristic of the second (especially in mythology)." Using this reasoning, you might look for an answer such as *wisdom* : *owl*, but none of the given choices has that kind of relationship. Another relationship between *pride* and *lion* is "a group of lions is called a pride"; therefore, the answer is (B) *pack* : *wolf*; "a group of wolves is called a pack."

Mathematics Example

If $16 \cdot 16 \cdot 16 = 8 \cdot 8 \cdot P$, then $P =$

(A) 4 (B) 8 (C) 32 (D) 48 (E) 64

This question car be solved by several methods. A time-consuming method would be to multiply the three 16s and then divide the result by the product of 8 and 8. A quicker approach would be to find what additional factors are needed on the right side of the equation to match those on the left side. These additional factors are two 2s and a 16, the product of which is 64. Yet another method involves solving for P as follows:

$$P = \frac{\overset{2}{\cancel{16}} \cdot \overset{2}{\cancel{16}} \cdot 16}{\cancel{8} \cdot \cancel{8}} = 2 \cdot 2 \cdot 16 = 64$$

The correct answer is (E).

•From *Taking the SAT* (College Entrance Examination Board, 1984).

APPENDIX B

Table B-1. Numbers of Actual Items Within Various Classifications for November SAT-Verbal Test Forms from 1970 to 1984

Item Type	Classification	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Sentence Completions	Aesthetics/philosophy	4	4	5	4	3	4	3	5	4	4	4	4	4	4	4
	World of practical affairs	5	6	5	5	5	5	5	4	4	4	4	5	4	4	4
	Science	5	5	5	5	4	3	4	3	4	4	4	3	4	4	4
	Human relationships	4	3	3	4	3	3	3	3	3	3	3	3	.3	3	3
	Total	18	18	18	18	15	15	15	15	15	15	15	15	15	15	15
Antonyms	Aesthetics/philosophy	4	4	5	5	6	6	6	6	5	6	6	6	7	4	6
	World of practical affairs	5	5	4	4	7	7	6	7	7	6	7	7	5	6	6
	Science	5	5	5	4	6	6	7	6	7	6	7	6	6	8	7
	Human relationships	4	4	4	5	6	6	6	6	6	7	5	6	7	7	6
	Total	18	18	18	18	25	25	25	25	25	25	25	25	25	25	25
Analogies	Aesthetics/philosophy	5	6	4	5	5	5	5	5	5	5	5	6	5	5	5
	World of practical affairs	5	5	6	5	6	5	6	5	5	6	5	5	5	5	6
	Science	5	5	5	5	4	5	5	5	5	4	5	5	5	5	5
	Human relationships	4	3	4	4	5	5	4	5	5	5	5	4	5	5	4
	Total	19	19	19	19	20	20	20	20	20	20	20	20	20	20	20
Reading Comprehension	Content															
	Narrative	5	5	5	5	5	5	5	5	5	4	5	5	3	5	3
	Biological science	5	10	5	5	5	5	5	5	4	5	4	3	5	3	4
	Physical science	5	0	5	5	0	0	0	0	3	3	5	4	3	4	5
	Argumentative	5	5	5	5	5	5	5	5	5	5	3	3	5	3	5
	Humanities	5	5	5	5	5	5	5	5	3	3	3	5	4	5	3
	Synthesis	5	5	5	5	0	0	0	0	0	0	0	0	0	0	0
	Social studies	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	Total	35	35	35	35	25	25	25	25	25	25	25	25	25	25	25
	Functional Skill															
	Main idea	4	6	5	6	6	4	5	4	4	5	3	3	5	4	5
	Supporting idea	7	9	10	8	5	5	4	5	7	7	8	6	6	7	7
	Inference	15	13	11	14	10	10	10	9	9	8	9	9	9	8	9
	Application	3	2	4	2	1	2	2	3	1	2	2	1	2	2	2
	Evaluation of logic	2	1	2	2	1	2	2	2	2	1	3	4	2	1	1
	Style and tone	4	4	3	3	2	2	2	2	2	2	0	2	1	3	1
	Total	35	35	35	35	25	25	25	25	25	25	25	25	25	25	25
Total Number of Items		90	90	90	90	85	85	85	85	85	85	85	85	85	85	85

BEST COPY AVAILABLE

Table B-2. Numbers of Actual Items Within Various Classifications for December SAT-Verbal Test Forms from 1970 to 1984

Item Type	Classification	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Sentence Completions	Aesthetics/philosophy	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
	World of practical affairs	5	5	5	5	4	4	4	4	4	4	4	4	4	4	4
	Science	5	5	5	5	4	4	4	4		4	4	4	4	4	4
	Human relationships	4	4	4	4	3	3	3	3		3	3	3	3	3	3
	Total	18	18	18	18	15	15	15	15	15	15	15	15	15	15	15
Antonyms	Aesthetics/philosophy	4	4	3	3	6	6	5	6	6	6	7	6	5	6	6
	World of practical affairs	5	5	7	6	6	7	7	6	6	6	5	6	7	6	6
	Science	5	5	5	5	7	7	7	7	6	7	4	7	6	7	7
	Human relationships	4	4	3	4	6	5	6	6	7	6	9	6	7	6	6
	Total	18	18	18	18	25	25	25	25	25	25	25	25	25	25	25
Analogies	Aesthetics/philosophy	5	5	5	5	5	5	5	5	5	4	5	5	4	5	5
	World of practical affairs	5	5	5	4	6	5	5	6	5	5	5	5	7	5	5
	Science	6	5	5	5	5	5	5	4	5	6	4	6	4	5	6
	Human relationships	3	4	4	5	4	5	5	5	5	5	6	4	5	5	4
	Total	19	19	19	19	20	20	20	20	20	20	20	20	20	20	20
Reading Comprehension	Content															
	Narrative	5	5	5	5	5	5	5	5	5	5	3	3	3	3	3
	Biological science	0	0	5	5	5	5	5	5	5	2	4	5	5	3	4
	Physical science	5	5	5	5	0	0	0	0	3	5	5	4	4	5	5
	Argumentative	10	10	10	5	5	5	5	5	3	4	5	3	3	5	3
	Humanities	5	5	5	5	5	5	5	5	5	4	5	5	5	4	5
	Synthesis	5	5	0	5	0	0	0	0	0	0	0	0	0	0	0
	Social studies	5	5	5	5	5	5	5	5	4	5	3	5	5	5	5
	Total	35	35	35	35	25	25	25	25	25	25	25	25	25	25	25
	Functional Skill															
	Main idea	5	6	2	8	5	4	3	3	1	4	3	6	4	3	4
	Supporting idea	4	7	7	7	6	5	10	6	7	5	8	2	5	5	8
	Inference	13	14	16	11	9	10	6	10	9	11	8	9	9	9	9
	Application	2	1	3	4	1	1	2	2	2	2	1	2	2	3	1
	Evaluation of logic	6	5	4	3	2	3	2	2	4	1	3	4	3	2	1
	Style and tone	5	2	3	2	2	2	2	2	2	2	2	2	2	3	2
	Total	35	35	35	35	25	25	25	25	25	25	25	25	25	25	25
Total Number of Items		90	90	90	90	85	85	85	85	85	85	85	85	85	85	85

Table B-3. Numbers of Actual Items Within Various Classifications for November SAT-Mathematical Test Forms from 1970 to 1984

Item Type	Classification	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Regular Math	Arithmetic	13	13	13	13	12	12	12	12	13	12	12	12	13	12	13
	Algebra	11	11	11	13	11	11	11	11	11	11	11	11	11	11	11
	Geometry	13	13	13	12	11	11	11	11	11	11	11	11	11	11	11
	Miscellaneous	5	5	5	4	6	6	6	6	5	6	6	6	5	6	5
	Total	42	42	42	42	40	40	40	40	40	40	40	40	40	40	40
Data Sufficiency	Arithmetic	4	4	4	4											
	Algebra	5	4	4	3											
	Geometry	5	6	6	8											
	Miscellaneous	4	4	4	3											
	Total	18	18	18	18											
Quantitative Comparisons	Arithmetic					6	6	7	6	6	6	6	6	6	6	6
	Algebra					6	6	5	6	6	6	6	6	6	6	6
	Geometry					6	6	6	6	5	5	5	5	6	6	5
	Miscellaneous					2	2	2	2	3	3	3	3	2	2	3
	Total					20	20	20	20	20	20	20	20	20	20	20
All	Setting															
	Concrete	17	17	14	12	13	14	16	10	11	14	16	17	19	14	16
	Abstract	43	43	46	48	47	46	44	50	49	46	44	43	41	46	44
	Ability:															
	Solving routine problems (Levels 0,1 and 2)	9	10	8	7	19	14	15	15	17	12	15	17	14	14	11
	Demonstrating comprehension of math ideas and concepts (Level 3)	21	23	22	22	23	22	26	23	24	29	25	24	25	29	31
	Applying "higher" mental processes to mathematics (Levels 4 and 5)	30[a]	27[a]	30[a]	31[a]	18	24	19	22	19	19	20	19	21	17	18
Total Number of Items		60	60	60	60	60	60	60	60	60	60	60	60	60	60	60

[a] All Data Sufficiency items were included in this category.

Table B-4. Numbers of Actual Items Within Various Classifications for December SAT-Mathematical Test Forms from 1970 to 1984

Item Type	Classification	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Regular Math	Arithmetic	13	13	13	13	12	12	12	12	13	12	13	13	13	12	12
	Algebra	11	11	11	12	11	11	11	11	11	11	10	11	11	11	11
	Geometry	13	13	13	12	11	12	11	11	11	11	11	11	11	11	11
	Miscellaneous	5	5	5	5	6	5	6	6	5	6	6	5	5	6	6
	Total	42	42	42	42	40	40	40	40	40	40	40	40	40	40	40
Data Sufficiency	Arithmetic	4	5	4	4											
	Algebra	5	4	4	5											
	Geometry	6	6	7	6											
	Miscellaneous	3	3	3	3											
	Total	18	18	18	18											
Quantitative Comparisons	Arithmetic					6	7	6	7	6	6	6	6	6	6	6
	Algebra					6	5	6	6	6	6	6	6	6	6	6
	Geometry					5	5	6	6	6	5	6	6	5	5	5
	Miscellaneous					2	3	2	1	2	3	2	2	3	3	3
	Total					20	20	20	20	20	20	20	20	20	20	20
All	Setting															
	Concrete	24	14	15	14	16	13	12	11	14	13	12	16	16	16	16
	Abstract	36	46	45	46	44	47	48	49	46	47	48	42	42	44	46
	Ability:															
	Solving routine problems (Levels 0,1, and 2)	9	7	8	9	17	20	12	11	16	18	10	15	12	13	15
	Demonstrating comprehension of math ideas and concepts (Level 3)	21	24	26	22	25	24	27	28	28	23	32	26	28	30	27
	Applying "higher" mental processes to mathematics (Levels 4 and 5)	30[a]	29[a]	26[a]	29[a]	18	16	21	21	16	19	18	17	20	17	18
Total Number of Items		60	60	60	60	60	60	60	60	60	60	60	60	60	60	60

[a] All Data Sufficiency items were included in this category.

Table B-5. Numbers of Items with Gender References of Particular Types in November SAT-Verbal Test Forms from 1970 to 1984[a]

Item Type	Gender Reference	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Sentence Completions	No Human Reference	1	4	4	2	4	3	4	6	5	4	6	5	3	2	3
	Human Reference															
	Male	6	4	4	6	3	5	6	2	4	5	2	3	3	4	1
	Female	0	0	1	1	2	0	0	0	3	2	0	1	1	2	3
	Both	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
	Neutral	10	3	9	9	3	7	5	7	6	3	6	6	8	7	8
	Generic he[b]	1	6	0	0	3	0	0	0	0	0	0	0	0	0	0
Analogies	No Human Reference	17	10	14	16	14	12	14	16	13	13	16	17	16	15	9
	Human Reference															
	Male	0	4	3	2	2	1	1	2	5	0	0	1	0	1	1
	Female	1	0	1	0	0	0	0	0	0	0	5	0	0	0	1
	Both	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
	Neutral	1	5	1	0	4	6	4	2	2	7	4	2	4	4	9
Reading Comprehension	No Human Reference	10	10	10	10	6	15	0	15	10	8	13	6	17	15	9
	Human Reference															
	Male	0	5	5	5	4	5	0	0	11	5	5	0	0	0	8
	Male with a reference to female(s)	5	0	0	0	0	0	0	5	0	3	0	0	0	0	0
	Reference to males[c]	5	5	10	5	10	0	10	0	4	5	3	14	5	5	6
	Female	0	0	0	0	0	0	0	0	0	0	0	0	3	5	0
	Both	0	0	5	5	0	0	10	0	0	4	0	5	0	0	0
	Neutral	5	0	0	0	0	0	0	0	0	0	5	0	0	0	0
	Generic he	10	15	5	10	5	5	5	5	0	0	0	0	0	0	0

[a] Based on data in Cruise and Kimmel (1989)

[b] An item was classified as containing a "generic he" only when it contained no other gender reference.

[c] An item in this category was not specifically about a male or males, yet it contained a reference to a male or males.

Table B-6. Numbers of Items with Gender References of Particular Types in December SAT-Verbal Test Forms from 1970 to 1984[a]

Item Type	Gender Reference	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Sentence Completions	No Human Reference	6	6	6	5	3	6	6	4	4	4	5	3	1	1	3
	Human Reference															
	Male	4	1	4	4	3	3	4	2	4	1	1	0	8	1	4
	Female	0	0	1	1	0	0	0	0	1	1	2	2	3	2	3
	Both	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
	Neutral	7	11	7	7	9	4	4	7	5	9	7	10	3	10	5
	Generic he[b]	1	0	0	1	0	2	1	1	0	0	0	0	0	0	0
Analogies	No Human Reference	13	14	14	12	15	13	17	14	15	18	15	15	13	16	15
	Human Reference															
	Male	3	2	2	3	1	2	2	1	1	1	2	1	0	1	1
	Female	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	Both	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0
	Neutral	3	3	2	3	3	5	1	4	4	1	3	4	6	2	4
Reading Comprehension	No Human Reference	0	5	10	0	5	5	5	5	8	10	9	9	12	8	0
	Human Reference															
	Male	5	0	5	5	5	0	10	0	5	5	3	0	0	4	3
	Male with a reference to female(s)	0	5	0	5	0	0	0	5	0	0	5	3	0	0	0
	Reference to males[c]	20	25	10	15	5	5	5	5	3	5	5	10	10	10	18
	Female	0	0	0	0	0	5	0	0	4	0	3	0	0	0	4
	Both	0	0	5	0	0	5	0	5	5	5	0	3	3	3	0
	Neutral	0	0	0	0	5	0	5	0	0	0	0	0	0	0	0
	Generic he	5	0	5	5	0	0	0	5	0	0	0	0	0	0	0
	Misc.	5	0	0	5	5	5	0	0	0	0	0	0	0	0	0

[a] Based on data in Cruise and Kimmel (1989)

[b] An item was classified as containing a "generic he" only when it contained no other gender reference.

[c] An item in this category was not specifically about a male or males, yet it contained a reference to a male or males.

Table B-7. Item-Difficulty (Equated-Delta) Distributions for November SAT-Verbal Test Forms from 1970 to 1984

Delta	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
≥ 18	-	-	-	-	-	-	-	-	1	-	-	-	1	-	-
17	2	2	1	2	3	1	1	1	-	-	2	-	-	-	-
16	4	6	5	8	7	5	6	4	3	4	3	2	1	2	3
15	5	6	5	7	7	9	12	6	10	12	10	13	9	9	3
14	14	9	8	10	12	11	5	8	6	9	8	9	12	9	18
13	10	11	7	8	5	8	8	10	13	7	9	7	9	9	11
12	10	8	13	11	3	4	4	7	6	6	6	7	7	10	5
11	13	13	11	10	8	8	7	8	4	5	5	8	11	13	9
10	10	11	11	11	9	7	8	7	13	7	5	6	7	5	5
9	7	7	9	5	8	6	8	10	6	12	8	9	10	5	10
8	5	6	10	6	9	9	8	5	11	6	8	8	5	10	7
7	8	7	3	7	5	9	8	12	9	12	10	7	5	6	6
6	2	3	5	1	8	3	6	4	-	1	6	5	3	2	4
≤ 5	-	1	2	4	1	5	4	3	3	3	4	4	5	4	4
No. of Items	90	90	90	90	85	85	85	85	85	84	84	85	85	84	85
Mean	11.9	11.8	11.5	11.7	11.5	11.4	11.3	11.3	11.4	11.4	11.1	11.2	11.5	11.4	11.3
SD	2.8	2.9	2.9	3.1	3.3	3.4	3.4	3.3	3.1	3.1	3.3	3.2	3.0	3.0	3.1

Table B-8. Item-Difficulty (Equated-Delta) Distributions for December SAT-Verbal Test Forms from 1970 to 1984

Delta	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
≥ 18	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	1	1	3	1	2	1	1	-	1	1	1	2	2	-	1
16	5	4	4	4	4	5	3	4	8	8	3	4	1	1	2
15	6	5	7	11	7	5	7	10	6	13	10	9	7	10	5
14	15	13	9	12	14	12	14	10	7	7	12	11	7	8	12
13	7	11	13	7	5	12	7	7	7	8	5	5	10	10	11
12	14	13	11	12	7	4	5	4	7	4	5	7	14	10	10
11	4	8	7	7	7	5	10	11	10	6	7	5	5	8	7
10	12	10	10	7	9	12	5	8	5	7	11	5	8	10	8
9	10	7	9	11	7	5	11	4	8	10	7	12	11	8	9
8	9	5	5	7	8	9	8	15	8	8	9	6	10	11	5
7	3	8	4	4	7	7	8	4	7	8	3	9	6	8	9
6	3	1	8	3	7	4	5	8	8	8	8	4	4	5	3
≤ 5	1	3	2	4	1	4	3	2	3	1	4	5	2	1	3
No of Items	90	90	90	90	85	85	85	85	85	85	85	84	85	84	85
Mean	11.9	11.8	11.8	11.7	11.5	11.4	11.4	11.3	11.4	11.5	11.3	11.3	11.3	11.2	11.4
SD	2.8	2.9	3.0	3.1	3.2	3.2	3.1	3.1	3.4	3.3	3.2	3.4	2.9	2.8	2.9

Table B-8. Item-Difficulty (Equated-Delta) Distributions for November SAT-Mathematical Test Forms from 1970 to 1984

Delta	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
≥ 18	-	-	2	3	4	3	1	2	-	1	2	2	-	4	4
17	4	3	4	4	3	3	3	3	2	6	2	5	7	1	5
16	3	6	4	6	6	1	6	3	7	6	4	4	4	5	4
15	6	6	6	3	2	3	5	8	4	3	6	4	5	6	2
14	8	4	2	2	4	7	5	3	5	2	4	4	3	4	6
13	3	2	5	7	6	2	2	4	3	3	6	2	8	1	2
12	3	8	9	5	5	6	8	6	6	7	5	6	2	6	6
11	12	8	7	9	8	9	5	9	11	10	6	6	8	4	10
10	7	6	8	7	5	7	5	8	8	8	8	6	9	12	8
9	7	9	5	6	4	5	11	7	5	7	6	12	4	8	4
8	3	3	4	3	3	7	1	4	5	2	5	1	6	7	5
7	2	3	2	3	4	4	7	3	2	1	1	4	2	-	3
6	1	-	2	1	3	2	-	2	2	3	1	3	1	1	1
≤ 5	-	1	-	1	1	1	1	-	1	1	1	-	1	1	-
No. of Items	60	60	60	60	60	60	60	60	60	60	59	59	60	60	60
Mean	12.3	12.2	12.4	12.6	12.5	11.8	12.1	12.2	12.1	12.3	12.2	12.1	12.3	12.1	12.6
SD	3.0	3.0	3.1	3.5	3.6	3.3	3.2	3.2	3.1	3.3	3.1	3.5	3.3	3.4	3.3

Table B-18. Item-Difficulty (Equated-Delta) Distributions for December SAT-Mathematical Test Forms from 1970 to 1984

Delta	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
≥ 18	2	2	1	-	1	3	1	3	3	-	1	1	1	-	3
17	2	3	2	3	1	3	1	5	4	5	5	-	3	3	3
16	4	4	9	3	6	5	7	8	4	4	5	6	6	8	3
15	5	5	1	7	6	5	-	2	5	4	3	4	8	3	7
14	7	4	7	7	6	3	8	2	2	3	2	9	1	4	2
13	5	7	5	6	5	6	4	2	3	10	9	5	5	2	5
12	5	8	-	11	2	6	8	3	6	3	2	9	6	6	5
11	8	5	13	3	6	2	7	8	10	9	8	8	2	8	7
10	6	6	9	3	13	5	8	9	6	8	7	7	11	8	7
9	7	10	6	9	5	12	9	8	7	6	8	4	7	5	7
8	4	2	4	5	3	2	3	6	3	2	6	2	6	5	4
7	4	3	-	-	3	5	2	1	3	2	2	1	2	5	2
6	1	1	1	2	3	3	-	2	2	2	1	3	2	1	1
≤ 5	-	-	2	1	-	-	2	-	1	2	1	1	-	1	3
No. of Items	60	60	60	60	60	59	60	59	59	60	60	60	60	59	59
Mean	12.3	12.3	12.3	12.5	12.1	11.8	12.0	12.4	12.2	12.2	12.2	12.4	12.1	12.0	12.2
SD	3.5	3.0	3.0	3.1	3.0	3.1	2.9	3.5	3.3	3.2	3.1	2.9	3.1	3.2	3.5

Table B-11. Scaled Scores Corresponding to Selected Raw Scores for November and December SAT-Verbal Test Forms from 1970 to 1984[a]

Year	Raw Score[b]					
	0	20/21	40/42	60/64	80/85	85/90
			November Administrations			
1970	230	360	500	640	770	800(810)
1971	220	350	490	630	770	800
1972	210	340	480	620	760	800
1973	230	360	500	630	770	800
1974	200(170)	320	460	610	760(750)	800(790)
1975	200	330	470	610	760(740)	800(770)
1976	200	340	470	600	750(740)	800(770)
1977	200(180)	320	460	600	760(740)	800(780)
1978	200(190)	330	460	600	750(730)	800(770)
1979	200(180)	320	460	600	750(740)	800(780)
1980	200(180)	310	440	580	750(710)	800(740)
1981	200(180)	320	450	590	750(730)	800(760)
1982	200(170)	320	460	580	730(710)	800(770)
1983	200(190)	340	460	590	730	800(780)
1984	200(180)	330	460	580	730	800(780)
			December Administrations			
1970	220	360	510	650	800	800(830)
1971	230	370	510	650	790	800(830)
1972	220	360	500	640	780	800(820)
1973	200(190)	340	490	640	790	800(820)
1974	200	340	480	620	760	800
1975	200(190)	330	470	610	750(740)	800(780)
1976	200(180)	320	460	600	750(740)	800(780)
1977	200	340	480	620	760	800(790)
1978	200(190)	320	460	600	750(740)	800(770)
1979	200(180)	320	460	600	760(750)	800(780)
1980	200(190)	330	460	600	760(730)	800(760)
1981	200(190)	330	460	590	750(720)	800(750)
1982	200(180)	320	450	580	740(720)	800(770)
1983	200	320	450	580	730	800(770)
1984	200	330	460	590	730	800(780)

[a] The scaled scores in parentheses are those that would have resulted without the application of "doglegs" to ensure that at least one raw score for each form converted to 800, and without the truncation of scores to the 200 to 800 scale.

[b] The scaled scores given for 1970 to 1973 correspond to raw scores 0, 21, 42, 64, 85, and 90; the scaled scores given for 1974 to 1984 correspond to raw scores 0, 20, 40, 60, 80, and 85.

Table B-12. Scaled Scores Corresponding to Selected Raw Scores for November and December SAT-Mathematical Test Forms from 1970 to 1984[a]

Year	Raw Score 0	15	20	30	40	45	60
November Administrations							
1970	280	410	450	540	630	680	800(810)
1971	250	390	440	530	620	670	800
1972	270	400	440	530	620	670	
1973	270	400	440	520	610	660(650)	800(780)
1974	270	400	450	540	620	670	800
1975	260	390	430	520	610(600)	660(650)	800(780)
1976	270	390	440	520	610	650	800(780)
1977	260	390	430	520	600	650	800(770)
1978	250	390	430	520	620	660	800
1979	260	390	430	520	600	650	800(780)
1980	250	390	430	520	610	660	800(790)
1981[b]	250	380	430	510	590	630	800(760)
1982	250	370	420	510	600	640	800(770)
1983	250	370	410	500	590	640	800(790)
1984	260	380	430	520	610	660	800(780)
December Administrations							
1970	270	410	460	550	640	690	800(830)
1971	260	400	440	530	630	670	800(810)
1972	280	410	460	550	640	680	800(810)
1973	270	410	450	540	630	670	800(810)
1974	260	400	440	530	620	670	800
1975	270	400	440	530	620(610)	670(660)	800(780)
1976	290	410	440	520	600	640	800(760)
1977	270	400	440	530	610	660(650)	800(780)
1978	260	390	440	530	610	660	800(790)
1979	270	400	450	530	620	660	800
1980	260	390	430	520	610	650	800(780)
1981	270	410	450	540	630	670	800
1982	280	380	410	500	600	650	800(790)
1983[b]	270	380	420	510	590	630	800(790)
1984	270	380	430	510	600	640	800(780)

[a] The scaled scores in parentheses are those that would have resulted without the application of "doglegs" to ensure that at least one raw score for each form converted to 800, and without the truncation of scores to the 200 to 800 scale.

[b] Only 39 items were scored; the scaled scores given correspond to raw scores 0, 15, 20, 30, 39, 44, and 59.

Table B-13. Numbers of SAT Scaled-Score Intervals with Scaled-Score Ranges of Particular Sizes[a,b]

Size of Scaled Score Range	March 1970 to April 1974	October 1974 to May 1978	October 1978 to December 1981	January 1982 to January 1985	Total
			SAT-Verbal		
40	8	0	4	6	18
30	9	4	10	10	33
20	0	12	2	0	14
10	0	0	1	0	1
0	1	2	1	2	6
			SAT-Mathematical		
50	2	0	0	2	4
40	4	0	0	6	10
30	4	6	6	4	20
20	2	6	6	0	14
10	0	0	0	0	0
0	1	1	1	1	4

[a] Based on the following number of test forms (forms with unscored items are not included):

	1970-74	1974-78	1978-81	1982-85
SAT-V	20	20	24	28
SAT-M	20	18	21	26

[b] Based on the scaled-score ranges in Tables 18 and 19 corresponding to 18 SAT-V scores and 13 SAT-M scores.

Table B-14. Equating Indices for November and December SAT-Verbal Equatings from 1970 to 1984

	First Equating				*Second Equating*				
Year	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.- Tot. r*	*Old Eq.- Tot. r*	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.- Tot. r*	*Old Eq.- Tot. r*	*Dif. Betw. Equat. Lines*
				November Administrations					
1970	.076	.979	.8780	.8727	-.088	.995	.8734	.8709	3.3
1971	-.068	.942	.8632	.8879	-.098	.927	.8647	.8777	1.7
1972	-.102	.965	.8701	.8756	-.105	.989	.8728	.8780	2.2
1973	-.140	1.018	.8658	.8557	.060	.983	.8545	.8600	1.6
1974	.085	1.022	.8808	Unav.	-.174	1.000	.8590	.8742	13.5
1975	.122	.998	.8766	.8738	-.114	1.024	.8636	.8571	13.8
1976	.008	.940	.8660	.8729	.004	.956	.8802	.8839	6.8
1977	-.005	.993	.8662	.8713	-.032	.942	.8725	.8764	5.3
1978	.121	1.028	.8745	.8696	-.051	.971	.8840	.8839	.7
1979	.032	1.019	.8682	.8721	.243	.999	.8626	.8618	.6
1980	-.055	.987	.8600	.8626	.218	.937	.8607	.8799	1.3
1981	.002	.991	.8574	.8555	.055	.928	.8522	.8773	.1
1982	-.005	.972	.8600	.8500	.353	.910	.8600	.8700	.6
1983	.010	.926	.8582	.8533	.097	.894	.8540	.8597	8.0
1984	.032	.974	.8500	.8605	.168	.933	.8639	.8790	7.9
				December Administrations					
1970	-.045	1.011	.8751	.8718	-.030	.972	.8702	.8795	6.8
1971	-.174	1.131	.8560	.8500	-.134	.966	.8675	.8751	.8
1972	-.290	.989	.8607	.8659	-.114	.980	.8528	.8752	1.6
1973	-.188	1.067	.8573	.8576	-.404	.957	.8683	.8814	16.6
1974	-.176	.947	.8393	.8471	-.062	.931	.8467	.8536	11.0
1975	.317	1.021	.8722	.8590	-.296	1.034	.8763	.8614	2.0
1976	-.192	.951	.8552	.8741	-.074	.990	.8612	.8722	10.7
1977	.008	1.023	.8610	.8520	-.150	.948	.8576	.8573	6.1
1978	.298	.932	.8626	.8743	-.101	.922	.8547	.8722	.9
1979	-.023	1.018	.8585	.8612	-.084	.849	.8553	.8778	.7
1980	-.065	.978	.8573	.8585	-.207	.921	.8627	.8671	14.8
1981	.005	.936	.8647	.8616	-.288	.944	.8533	.8661	6.2
1982	.027	.999	.8504	.8470	-.175	.895	.8552	.8671	2.6
1983	-.012	.974	.8587	.8528	-.292	.968	.8592	.8580	.3
1984	-.054	.957	.8677	.8664	-.262	.941	.8607	.8579	10.7
Boundary:									
Top	.0020	1.0000	.8879	.8879	.0020	1.0000	.8879	.8879	.07
2 - 3	.1360	1.0561	.8735	.8735	.1360	.9469	.8735	.8735	5.58
1 - 2	.2700	1.1153	.8573	.8573	.2700	.8966	.8573	.8573	11.09
Bottom	.4040	1.1779	.8393	.8393	.4040	.8490	.8393	.8393	16.60

Note: The boundaries listed for variance ratios differ depending on whether the ratios are less than one or greater than one. The boundaries listed for the first equatings are for ratios greater than or equal to one; those listed for the second equatings are for ratios less than or equal to one.

Table B-15. Equating Indices for November and December SAT-Mathematical Equatings from 1970 to 1984

	First Equating				*Second Equating*				
Year	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.-Tot. r*	*Old Eq.-Tot. r*	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.-Tot. r*	*Old Eq.-Tot. r*	*Dif. Betw. Equat. Lines*
				November Administrations					
1970	.073	.880	.8524	.8603	-.122	.947	.8453	.8546	.5
1971	-.073	.958	.8351	.8495	-.111	.951	.8265	.8449	6.1
1972	-.094	1.011	.8407	.8540	-.040	1.025	.8442	.8524	6.5
1973	-.098	.958	.8583	.8685	.043	1.124	.8668	.8619	1.6
1974	.054	.930	.8329	.8562	-.131	1.028	.8512	.8619	3.7
1975	.134	1.020	.8585	.8478	-.038	1.100	.8740	.8506	7.7
1976	.064	1.046	.8755	.8617	-.011	1.017	.8733	.8786	8.3
1977	.0004	.980	.8595	.8696	-.042	1.024	.8631	.8533	7.1
1978	.040	.994	.8283	.8405	-.104	.997	.8616	.8786	4.7
1979	.039	1.022	.8622	.8600	.221	1.002	.8535	.8559	9.1
1980	-.051	.992	.8376	.8535	.233	.907	.8613	.8767	7.6
1981	-.004	.923	.8249	.8335	.082	.980	.8487	.8784	12.8
1982	-.002	.961	.8500	.8600	.257	.924	.8400	.8600	3.0
1983	.020	.983	.8332	.8370	.035	.876	.8483	.8551	.6
1984	.086	1.082	.8454	.8477	.159	.951	.8534	.8684	10.5
				December Administrations					
1970	.002	.974	.8727	.8760	-.057	1.005	.8514	.8492	3.6
1971	-.125	1.129	.8535	.8528	.019	.969	.8463	.8602	.8
1972	-.080	1.142	.8516	.8360	-.124	.970	.8503	.8727	1.6
1973	-.173	1.101	.8654	.8545	-.354	1.115	.8673	.8468	15.5
1974	-.023	.987	.8517	.8547	-.182	1.006	.8326	.8362	23.1
1975	-.262	1.038	.8534	.8329	-.235	1.021	.8524	.8615	2.6
1976	-.125	.979	.8715	.8654	.012	1.052	.8605	.8536	7.5
1977	-.035	1.002	.8680	.8691	-.147	.903	.8619	.8654	4.3
1978	-.291	.960	.8355	.8469	-.056	.968	.8345	.8534	1.4
1979	.007	.998	.8612	.8637	-.053	.923	.8543	.8749	1.7
1980	.012	.994	.8583	.8612	-.151	.965	.8413	.8697	4.4
1981	-.042	.959	.8362	.8395	-.247	1.012	.8398	.8478	2.2
1982	-.008	.957	.8372	.8395	-.081	.936	.8421	.8638	8.3
1983	.033	1.018	.8481	.8370	-.251	1.023	.8751	.8695	4.9
1984	.034	.969	.8635	.8611	-.210	.875	.8204	.8386	3.2
Boundary:									
Top	.0004	1.0020	.8786	.8786	.0004	.9980	.8786	.8786	.50
2 - 3	.1183	1.0469	.8615	.8615	.1183	.9552	.8615	.8615	8.03
1 - 2	.2361	1.0938	.8422	.8422	.2361	.9142	.8422	.8422	15.57
Bottom	.3540	1.1429	.8204	.8204	.3540	.8750	.8204	.8204	23.10

Note: The boundaries listed for variance ratios differ depending on whether the ratios are less than one or greater thɪn one. The boundaries listed for the first equatings are for ratios greater than or equal to one; those listed for the second equatings are for ratios less than or equal to one.

Table B-16. Summary Equating Indices for November and December SAT-Verbal Equatings from 1970 to 1984

	First Equating				*Second Equating*						
Year	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.-Tot. r*	*Old Eq.-Tot. r*	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.-Tot. r*	*Old Eq.-Tot. r*	*Dif. Betw. Equat. Lines*	*Over-All Comp.*	*Period Avg.*
					November Administrations						
1970	3	3	3	2	3	3	2	2	3	2.9	
1971	3	2	2	3	3	2	2	3	3	2.8	
1972	3	3	2	3	3	3	2	3	3	2.9	
1973	2	3	2	1	3	3	1	2	3	2.6	2.78
1974	3	3	3	Unav.	2	3	2	3	1	2.1	
1975	3	3	3	3	3	3	2	1	1	2.2	
1976	3	2	2	2	3	3	3	3	2	2.5	
1977	3	3	2	2	3	2	2	3	3	2.8	2.41
1978	3	3	3	2	3	3	3	3	3	3.0	
1979	3	3	2	2	2	3	2	2	3	2.7	
1980	3	3	2	2	2	2	2	3	3	2.6	
1981	3	3	2	1	3	2	1	3	3	2.7	2.74
1982	3	3	2	1	1	2	2	2	3	2.4	
1983	3	2	2	1	3	1	1	2	2	2.2	
1984	3	3	1	2	2	2	2	3	2	2.3	2.26
					December Administrations						
1970	3	3	3	2	3	3	2	3	2	2.6	
1971	2	1	1	1	3	3	2	3	3	2.5	
1972	1	3	2	2	3	3	1	3	3	2.5	
1973	2	2	1	2	1	3	2	3	1	1.6	2.28
1974	2	3	1	1	3	2	1	1	2	2.1	
1975	1	3	2	2	1	3	3	2	3	2.2	
1976	2	3	1	3	3	3	2	2	2	2.3	
1977	3	3	2	1	2	3	2	1	2	2.3	2.22
1978	1	2	2	3	3	2	1	2	3	2.3	
1979	3	3	2	2	3	1	1	3	3	2.7	
1980	3	3	1	2	2	2	2	2	1	1.9	
1981	3	2	2	2	1	2	1	2	2	2.0	2.21
1982	3	3	1	1	2	1	1	2	3	2.4	
1983	3	3	2	1	1	3	2	2	3	2.5	
1984	3	3	2	2	2	2	2	2	2	2.3	2.36
Boundary:											
Top	.0020	1.0000	.8879	.8879	.0020	1.0000	.8879	.8879	.07		
2 - 3	.1360	1.0561	.8735	.8735	.1360	.9469	.8735	.8735	5.58		
1 - 2	.2700	1.1153	.8573	.8573	.2700	.8966	.8573	.8573	11.09		
Bottom	.4040	1.1779	.8393	.8393	.4040	.8490	.8393	.8393	16.60		

Note: The boundaries listed for variance ratios differ depending on whether the ratios are less than one or greater than one. The boundaries listed for the first equatings are for ratios greater than or equal to one; those listed for the second equatings are for ratios less than or equal to one.

Table B-17. Summary Equating Indices for November and December SAT-Mathematical Equatings from 1970 to 1984

	First Equating				*Second Equating*						
Year	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.-Tot. r*	*Old Eq.-Tot. r*	*Std. Mean Dif.*	*Var. Ratio*	*New Eq.-Tot. r*	*Old Eq.-Tot. r*	*Dif. Betw. Equat. Lines*	*Over-All Comp.*	*Period Avg.*
					November Administrations						
1970	3	1	2	2	2	2	2	2	3	2.4	
1971	3	3	1	2	3	2	1	2	3	2.7	
1972	3	3	1	2	3	3	2	2	3	2.8	
1973	3	3	2	3	3	1	3	3	3	2.8	
1974	3	2	1	3	2	3	2	3	3	2.6	2.67
1975	2	3	2	2	3	1	3	2	3	2.5	
1976	3	3	3	3	3	3	3	3	2	2.7	
1977	3	3	2	3	3	3	3	2	3	2.9	
1978	3	3	1	1	3	3	3	3	3	2.8	2.69
1979	3	3	3	2	2	3	2	2	2	2.4	
1980	3	3	1	2	2	1	2	3	3	2.5	
1981	3	2	1	1	3	3	2	3	2	2.4	
1982	3	3	2	2	1	2	1	2	3	2.4	2.52
1983	3	3	1	1	3	1	2	2	3	2.6	
1984	3	2	2	2	2	2	2	3	2	2.2	2.39
					December Administrations						
1970	3	3	3	3	3	3	2	2	3	2.9	
1971	2	1	2	2	3	3	2	2	3	2.5	
1972	3	1	2	1	2	3	2	3	3	2.5	
1973	2	1	3	2	1	1	3	2	2	1.8	
1974	3	3	2	2	2	3	1	1	1	1.9	2.42
1975	1	3	2	1	2	3	2	3	3	2.3	
1976	2	3	3	3	3	2	2	2	3	2.7	
1977	3	3	3	3	2	1	3	3	3	2.7	
1978	1	3	1	2	3	3	1	2	3	2.4	2.40
1979	3	3	2	3	3	2	2	3	3	2.8	
1980	3	3	2	2	2	3	1	3	3	2.7	
1981	3	3	1	1	1	3	1	2	3	2.4	
1982	3	3	1	1	3	2	1	3	2	2.3	2.57
1983	3	3	2	1	1	3	3	3	3	2.5	
1984	3	3	3	2	2	1	1	1	3	2.5	2.44
Boundary:											
Top	.0004	1.0020	.8786	.8786	.0004	.9980	.8786	.8786	.50		
2 - 3	.1183	1.0469	.8615	.8615	.1183	.9552	.8615	.8615	8.03		
1 - 2	.2361	1.0938	.8422	.8422	.2361	.9142	.8422	.8422	15.57		
Bottom	.3540	1.1429	.8204	.8204	.3540	.8750	.8204	.8204	23.10		

Note: The boundaries listed for variance ratios differ depending on whether the ratios are less than one or greater than one. The boundaries listed for the first equatings are for ratios greater than or equal to one; those listed for the second equatings are for ratios less than or equal to one.