CollegeBoard

Research Report
No. 2004-2

A Simulation Study to Explore Configuring the New SAT® Critical Reading Section Without Analogy Items

Jinghua Liu
Miriam Feigenbaum
Linda Cook

connect to college success
www.collegeboard.com

College Board Research Report No. 2004-2
ETS RR-04-01

A Simulation Study to Explore Configuring the New SAT® Critical Reading Section Without Analogy Items

Jinghua Liu, Miriam Feigenbaum, and Linda Cook

College Entrance Examination Board, New York, 2004

Jinghua Liu is measurement statistician at Educational Testing Service.

Miriam Feigenbaum is principal statistical associate level II at Educational Testing Service.

Linda Cook is principal research scientist level II at Educational Testing Service.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board: Expanding College Opportunity

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,500 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #030481025) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Contents

Abstract

This study explored possible configurations of the new SAT® critical reading section without analogy items. The item pool contained items from SAT verbal (SAT–V) sections of 14 previously administered SAT tests, calibrated using the three-parameter logistic IRT model. Multiple versions of several prototypes that do not contain analogy items were assembled. Item statistics and test statistics for the simulated forms were compared to the average of 13 forms of the SAT–V. These statistics included: IRT scaled score reliability, scaled score standard error of measurement, conditional scaled score standard error of measurement, *r*-biserial, and equated deltas. The results indicated that it is possible to maintain measurement precision for the new SAT critical reading section without analogy items, but it may be necessary to modify the distribution of item difficulty in order to obtain adequate precision at the ends of the score scale.

Key words: SAT verbal section (SAT–V), new critical reading section, data mining, analogy

Introduction

The SAT Reasoning Test™ (hereinafter called the SAT) is an objective and standardized test that measures verbal and mathematical reasoning abilities that students develop over time, both in and out of school. The current verbal portion of the test (SAT–V) measures verbal reasoning abilities, with emphasis on critical reasoning and vocabulary abilities in the context of reading passage, analogy, and sentence completion questions. The SAT–V includes 78 items: 19 *Analogy* (AN) items, 19 *Sentence Completion* (SC) items, and 40 *Critical Reading* (CR) items. Each of the three item types allows measurement of vocabulary knowledge, and all include a range of words of various levels of difficulty.

In order to strengthen the alignment of the SAT to curriculum and instructional practices in high schools and colleges, the College Board will be making substantial changes to the SAT. For the verbal portion, the upcoming changes include eliminating AN items, adding paragraph-length critical reading passages, and changing the name of the test section from verbal to critical reading. The new SAT critical reading section "measures knowledge of genre, cause and effect, rhetorical devices, and comparative arguments and ability to recognize relationships among parts of a text" (College Board, 2002). Vocabulary knowledge will continue to be measured, but through the use of "vocabulary in context" questions, based either on reading passages or independent sentences.

This study explored the configuration of the new SAT critical reading section without AN items. The study was carried out as a simulation study using data from past administrations of the SAT–V. Simulated forms with a reduced number of or no AN items were assembled. Item and test statistics for the simulated forms were analyzed and compared to the current forms. An underlying assumption in the reconfiguration was that the overall difficulty of the SAT–V and the score reporting scale for the SAT–V would not be changed. Consequently, an important constraint for the study was the maintenance, as closely as possible, of the current test specifications including delta distributions. The current specifications were established, using IRT methods and were endorsed by the SAT Committee when the SAT–V and mathematical (SAT–M) sections were revised in 1994. In this study, we worked carefully to maintain the overall difficulty level of the SAT–V and to mirror, as much as possible, the psychometric characteristics of the SAT–V.

This study represents an application of IRT to explore the implications of revising an existing test section, such as the SAT–V. IRT provides a powerful data simulation tool to evaluate the impact of revising a test section, as long as item responses exist for all of the items involved in the revision. Previous experience using the three-parameter logistic model with the SAT–V and SAT–M indicates that this model fits the data well. (See Cook and Petersen, 1987; Petersen, Cook, and Stocking, 1983.) Consequently, it was possible to use this model to simulate the impact on test and item statistics, such as test reliability and conditional standard errors of measurement, to evaluate the efficacy of various test configurations that were simulated without the analogy item type.

There were three phases in this study. During Phase 1, multiple versions of four prototypes without or with a reduced number of AN items were assembled and analyzed. However, none of the prototypes produced test simulations with measurement errors that are as small as those produced by the SAT–V for scores below 300 or above 700. As a result, a second phase of the study was carried out. Two of the most promising prototypes were selected and revised, and additional versions of each of these two prototypes were simulated and analyzed. Further investigations were explored during Phase 3. In an effort to better understand and compare the measurement errors resulting from different item types, three hypothetical tests were formed using IRT item statistics: an all-AN item test, an all-SC item test, and an all-CR item test. The conditional standard error of measurement of each hypothetical test was computed and compared.

Phase 1

Method

Design of Prototypes

Table 1 provides the configuration of the current verbal section (SAT–V) as well as the four prototypes for the new critical reading section evaluated for the study. The prototypes were created by ETS experts given consideration of face validity, speededness, alignment with current test, etc., and consultation with the SAT Test Development Committee. All prototypes, with the exception of Prototype C, were assembled without AN items. Prototype C, which contained 10 AN items, was assembled to provide additional baseline data for the study and also as a possible alternative in the event that it was found that the omission of AN items seriously impacted the ability to produce a viable replacement for the current verbal section. Prototypes A and B represented increasingly heavier reliance on a reading construct. Prototype A contained approximately 56 percent CR items, as compared to 51 percent CR items that appear in the SAT–V. Prototype B contained approximately 71 percent CR items. Prototype D also contained a high percentage of CR items (approximately 72 percent), and included a simulated item type, *Discrete Reading* (DR) items that do not appear in the SAT–V. DR items each have a stimulus of 60–80 words (two or three sentences) followed by a single multiple-choice question. When the study was conducted, there was no information available on how these items would function. ETS experts estimated the item statistics based on their experiences and the characteristics of the current critical reading items. These DR items were pretested later in a regular SAT administration. The item statistics turned out to be very close to the estimations.

All prototypes were shorter than the SAT–V because the administration time for the new critical reading section will be reduced to 70 minutes from the current 75 minutes. In addition, the amount of time needed to answer different types of items is different. It was estimated that the average time to answer each item type is 0.5 min/AN, 0.7 min/SC, and 1.0–1.2 min/CR depending on the length of the passage (Bridgeman, Cahalan, and Cline, 2003). As can be seen, the AN item type is least time consuming. When the testing time is shorter and the least time-consuming items are removed, it is necessary to shorten the test in order to ensure that the sections are not speeded.

TABLE 1

Configuration of the Current SAT® Verbal Section (SAT–V) and Prototypes (Number of Items by Item Types)

		Prototype			
	Current	*A*	*B*	*C*	*D*
Analogy	19	-	-	10	-
Sentence Completion	19	32	19	25	19
Critical Reading	40	40	46	40	40
Discrete Reading	-	-	-	-	8
Total	78	72	65	75	67

Note: Discrete Reading items each have a "passage" of 60–80 words (two or three sentences). This item type does not exist in the current configuration of the test section (SAT–V).

Item Pool Construction

The item pool was formed by linking the item parameter estimates from the Item Response Theory (IRT) calibrations of operational and equating SAT items that had been administered over the past several years. Fourteen SAT–V operational forms and several SAT–V equating tests from previous administrations were calibrated using the three-parameter logistic IRT model, and the resulting parameter estimates were placed on the same scale by linking them back to the same base form. The linking procedure used to place item parameter estimates from the 14 tests on the same scale was developed by Stocking and Lord (1983), and was found by Petersen, Cook, and Stocking (1983) to work well with SAT data. These calibrated items formed the SAT–V item pool containing more than 1,300 items, which provided the basis for the simulation forms.

Assembly of Simulated Forms

Automated item selection (AIS). To construct tests, items are selected and assembled into intact test forms. Item selection is usually subject to various rules to constrain the selection of items for test forms. These rules are called test specifications. Test specifications for the SAT can be classified into several categories: content constraints, statistical constraints, item sensitivity, and item overlap. When constructing tests, test developers provide a set of constraints to a computer, and then evaluate the results of the selected items.

The SAT Program currently employs a test creation software that uses the automated item selection (AIS) algorithm to assemble tests. This method requires a variety of constraints with different weights, and these constraints will be used as rules to select items. This model attempts to satisfy the target test properties by minimizing the aggregate failures, and attempts to provide some control over allowable failures by weighting each constraint (Stocking and Swanson, 1992). When allowable failures happen, lower-weighted constraints will be violated first. Consequently, by using AIS, the quality of the assembled test is usually assured. When building multiple versions of

the tests, AIS is capable of developing unique tests without any or very low item overlap.

It was decided to use AIS to assemble prototypes for this study. All of the constraints including content constraints, item sensitivity, and item overlap remained the same as the SAT–V. Therefore, the modified prototypes could be assembled exactly the same way as the SAT–V is assembled.

Setting statistical specifications. Statistical specifications provide the guide for building tests that discriminate effectively at the ability levels where discrimination is most needed. SAT statistical specifications call for specific numbers of items across a range of intervals on the item difficulty scale. At ETS, the index of item difficulty used for the SAT Program is the delta statistic. The delta index is based on the percent of test-takers who attempt to answer the item and who answer the item correctly (i.e., *p*-value), where 1 minus *p*-value are converted to a normalized *z*-score and transformed to a scale with a mean of 13 and a standard deviation of 4. A higher delta value represents a harder item.

This conversion of *p*-values provides raw delta values that reflect the difficulty of the items taken by particular examinees from a particular administration. This measure of item difficulty then must be adjusted to correct for differences in the ability of different test-taking populations. Delta equating is a statistical procedure used to convert raw delta values to equated delta values. This procedure involves administering some old items with known equated delta values, along with new items. Each old item now has two difficulty measures: the equated delta, which is on the scale, and the observed delta from the current group of examinees. The linear relationship between the pairs of observed and equated deltas on the old items is used to determine scaled values for each of the new items. Delta equating is essential because the groups taking a particular test may differ substantially in ability from one administration to another. Through delta equating, the difficulty of items taken by different groups can be expressed on a single scale so they can be more appropriately compared. The delta values discussed in this paper are equated deltas.

As mentioned previously, SAT test specifications have historically been set using equated delta distributions. This practice was continued when the test was revised in 1994. Consequently, the test is assembled using classical test theory statistics (equated deltas and biserial correlation coefficients). All test assembly software developed for the SAT operates using target distributions of these classical test theory statistics.

The delta distribution for the SAT–V is shown in Table 2. As can be seen, it is a unimodel distribution with more middle difficulty items and fewer very easy or very difficult items. Rather than proposing a new delta distribution, the delta distributions were obtained for each prototype by proportionally reducing the number of items at each delta level to reflect the reduced total number of items in the prototypes. Since the same proportion of items were maintained at each delta level, the mean and standard deviation of each prototype were very close to the specified equated delta mean and standard deviation of the SAT–V. As mentioned previously, an important constraint of this study was the maintenance of the overall difficulty level of the current SAT, so matching the prototype specifications to the current test specifications was an important step in the study.

TABLE 2

Specified Delta Distributions for the SAT–V and Prototypes

Delta Level (Specified)	*Current*	*Prototype (10 forms/prototype)*			
		A	*B*	*C*	*D*
>=19	-	-	-	-	-
18	-	-	-	-	-
17	-	-	-	-	-
16	1	1	1	1 (1–2)	1 (0–1)
15	4	3	3	4 (3–4)	3 (2–3)
14	6	6	5	6 (5–6)	5 (4–6)
13	8	8 (7–8)	7 (6–7)	8 (7–8)	7 (6–8)
12	12	11 (11–12)	10 (10–12)	11 (11–12)	10 (8–13)
11	14	13 (13–14)	11 (10–14)	13 (13–14)	13 (11–15)
10	12	11	10 (9–10)	11 (11–12)	10 (9–12)
9	9	8	8 (7–9)	9	8 (7–9)
8	7	6 (5–6)	6 (5–6)	7 (7–8)	6 (5–7)
7	4	4	3	4 (4–5)	3 (3–4)
6	1	1	1	1 (1–2)	1 (0–1)
<=5.9	-	- (0–1)	- (0–1)	- (0–1)	- (0–1)
Number of Items	78	72	65	75	67
Mean	11.4	11.4	11.4	11.4	11.4
S.D.	2.2	2.2	2.2	2.3	2.2

Note: The numbers in the parentheses are the actual ranges used in the prototypes.

Assembly of simulation forms. Once statistical specifications for all prototypes were set, test assembly software, AIS, was used to assemble simulated test forms from the item pool. Ten versions of each of the four prototypes were assembled for a total of 40 experimental test versions. Each of the 10 versions under a prototype was a unique test without any item overlap.

The prototypes were evaluated in terms of item statistics, test statistics, and nonpsychometric criteria. The results were compared to criteria established based on selected statistics from 13 SAT forms, administered from March 1999 to May 2001.

Analyses of Prototypes: Item Statistics

Item statistics are statistical descriptions of how a particular item functions in a test. Typically analyses provide information about the difficulty of the item and the ability of the item to discriminate among the examinees.

As described previously, equated delta is the difficulty index reported in this study. The other item statistic evaluated in this study is item discrimination power. Each item in a test should be able to distinguish between higher ability and lower ability examinees with respect to the trait being measured. The degree to which an item can discriminate between higher ability and lower ability examinees is known as its power of discrimination. There are a number of methods of assessing the discriminating power of a test item. The one currently used at ETS is the *r*-biserial, which measures the strength of relationship between a dichotomous variable (item right versus item wrong) and a criterion variable that is continuous (a total test score along the score scale with many possible values).

For each of the prototypes, the item statistics (equated deltas and *r*-biserial) were produced by averaging the item statistics (10 forms/prototype) when the individual items were administered as part of the operational test forms.

Analyses of Prototypes: Test Statistics

Test statistics provide information on precision of measurement. The estimates of test statistics reported in this paper are IRT scaled score reliability, IRT scaled score standard error of measurement (SEM), and IRT scaled score conditional standard error of measurement (CSEM).

The psychometric properties of the prototypes were evaluated by comparing the IRT scaled score reliability, SEM, and CSEM to these same statistics obtained from the SAT–V. The statistics were obtained using algorithms described by Dorans (1984). Dorans described the computation of CSEMs that can be combined with ability distributions to produce IRT-based estimates of the unconditional standard error of measurement as well as a reliability coefficient. The formulas developed by Dorans are described below.

Dorans (1984) employed the three-parameter IRT model as follows,

$$(1)\quad p_i\left(\theta_j\right) = c_i + \frac{1-c_i}{1+\exp\left[-1.7a_i\right]\cdot\left(\theta-b_i\right)}$$

where $P_i(\Theta_j)$ is the probability of a correct response at a given ability level, given three item properties: discrimination (a_i), difficulty (b_i), and guessing (c_i).

The CSEM for a given number right score, based on the binomial model, is

$$(2)\quad CSEM\left(RS_j\middle|\theta_j\right) = \sqrt{\sum_{i=1}^{K_T} P_i\left(\theta_j\right)Q_i\left(\theta_j\right)},$$

where $Q_i(\Theta_j) = 1\text{-}P_i(\Theta_j)$, and K_T is the total number of items on the test. When formula scoring is used, the CSEM may be computed by,

$$(3)\quad CSEM\left(FS_j\middle|\theta_j\right) = \sqrt{\sum_{i=1}^{K_T}\left(\frac{k_i}{k_i-1}\cdot CSEM\left(RS_j\middle|\theta_j\right)\right)^2},$$

where K_i is the number of alternatives associated with item *i*.

An overall SEM is calculated based on CSEMs and the number of test-takers,

$$(4)\quad SEM = \sqrt{\frac{\sum_j\left(CSEM_j^{\,2}\cdot N_j\right)}{N_T}},$$

where N_j is the number of examinees obtaining score *j* in the analysis sample, and N_T is the total number of examinees in the analysis sample.

IRT scaled score reliability is calculated from SEM.

$$(5)\quad reliability = 1-\frac{SEM^2}{\sigma^2},$$

where σ^2 is the variance of the scores.

The curves for the CSEMs were produced as part of the IRT equating analyses available using the GENASYS software[1]. Scores on all prototype forms were equated to scores on the same base form administered in March 2001. The criterion forms used for graphical displays of CSEMs are the 13 SAT–V forms that were previously mentioned.

The frequency distributions required to compute these reliability estimates were constructed using the distribution of scores obtained from the base form administration and the equating relationship between the prototypes and the base form. Figure 1 shows the frequency distribution for a hypothetical base form, to which total scores on all simulated forms are equated as a result of having item parameter estimates on the same scale as the base form. In this example, we equate a new simulated form to the base form. Suppose we want to know the frequency at a score of 56 on the new form.

[1] GENASYS (Generalized Analysis System) is a comprehensive statistical analysis system that combines current computer technology with modern psychometric practices to perform various statistical analyses.

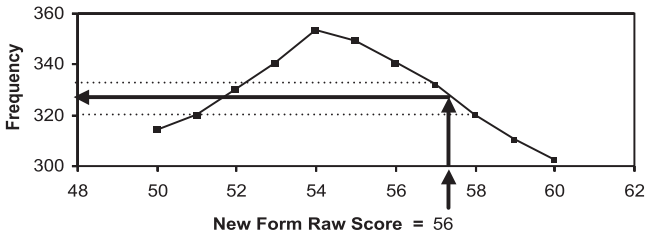

Figure 1. Interpolation of frequency distribution for the equated new form from the base form.

The corresponding equated raw score (i.e., the equivalent score on the base form) is 57.7. We use the base form frequency distribution to estimate the frequency at this point. The graph shows that 332 people received a score of 57 on the base form, whereas 320 people received a score of 58. Interpolating from the graph, we would estimate that 328 people would receive a score of 56 on the new form.

Establishing Psychometric Criteria

The criteria used for the evaluation of the prototypes were constructed in several ways. The mean and standard deviation of equated deltas, *r*-biserials, SEMs, and the reliability coefficients were computed by averaging values taken from the analyses of 13 SAT–V forms. Criteria for the CSEMs were the average value at each scaled score level of these 13 forms.

Establishing Nonpsychometric Criteria

In addition to psychometric analyses, ETS content experts were asked to evaluate each of the prototypes according to the following nonstatistical criteria: face validity; educational relevance; ease of development; ease of configuring the SAT–V into separately timed sections; the cost of transitioning to the new critical reading section; the ongoing operational costs once the transition period was over; the ability to sustain subscores (should they be desired at some point in time); and the ease of aligning the PSAT/NMSQT® with the recommended changes to the SAT–V.

Results

Item Statistics—Equated Deltas and *r*-biserial

Table 3 provides information about the item statistics for the prototypes. It can be seen that the mean and standard deviation of equated deltas and the mean and standard deviation of *r*-biserial obtained for the prototypes are very similar to those for the criteria.

TABLE 3

Summary of Item Statistics for the SAT–V and Prototypes

	Current	*Prototype*			
		A	*B*	*C*	*D*
Number of Items	78	72	65	75	67
Equated Delta					
Mean	11.4	11.4	11.4	11.3	11.3
S.D.	2.2	2.2	2.2	2.3	2.2
r-biserial					
Mean	0.51	0.53	0.51	0.52	0.52
S.D.	0.10	0.10	0.09	0.09	0.10

Note: Current criteria are based on 13 SAT–V forms.

Test Statistics—CSEMs, SEMs, and Reliabilities

Plots of IRT scaled score CSEM. Plots of IRT scaled score conditional standard errors of measurement can be found in Figures 2 through 9. These plots show the average CSEM values for the multiple versions of each prototype compared to the average CSEM values for the criterion obtained by averaging across 13 SAT–V forms, as well as the CSEM values for the 10 versions of each prototype compared to the CSEM values for 13 SAT–V forms.

An examination of the average CSEM of Prototype A compared to the criterion, found in Figure 2, shows that slightly greater precision of measurement was gained for scores between about 300 and 700, where the majority of the scores are located. However, some measurement power was lost below 300 and above 700, where the CSEMs for the prototype were larger than those for the criterion. In addition, the CSEM values for all of the 10 versions under Prototype A were compared to the CSEM values for all 13 SAT–V forms that were used as criteria (see Figure 3). The results indicated that although there was some variation across individual forms, the trend was the same: Most of the simulated forms had larger CSEMs than the 13 criterion forms for scaled scores below about 300 and above approximately 700.

Figures 4 and 5 show plots of CSEMs for Prototype B, average and multiple versions, respectively. The average of Prototype B appeared to result in slightly larger CSEMs throughout the mid-portion of the score range and larger CSEMs over the ends of the score range when compared to the criterion. The 10 individual forms under Prototype B followed a similar pattern.

Plots for the average of the 10 versions of Prototype C are found in Figure 6, and plots for multiple versions of Prototype C are found in Figure 7. Prototype C versions appeared to produce slightly smaller CSEMs than the criterion throughout the mid-portion

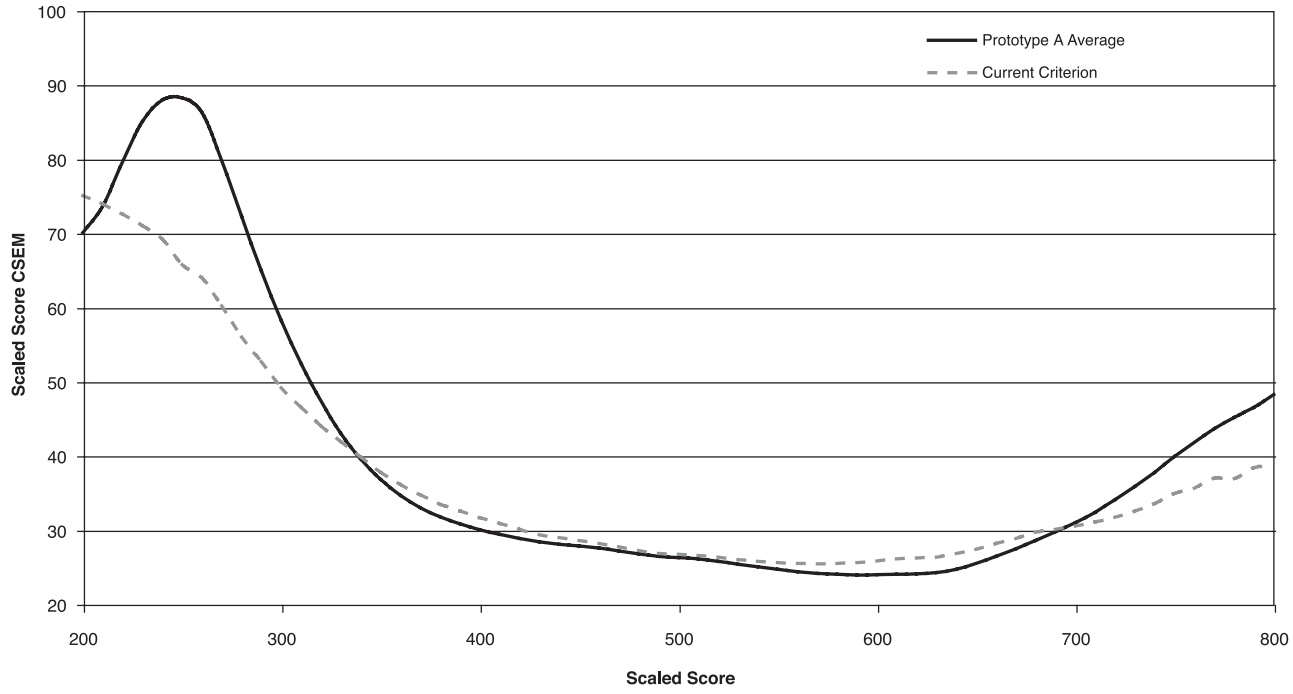

Figure 2. The average scaled score CSEMs: Prototype A compared to the SAT–V.

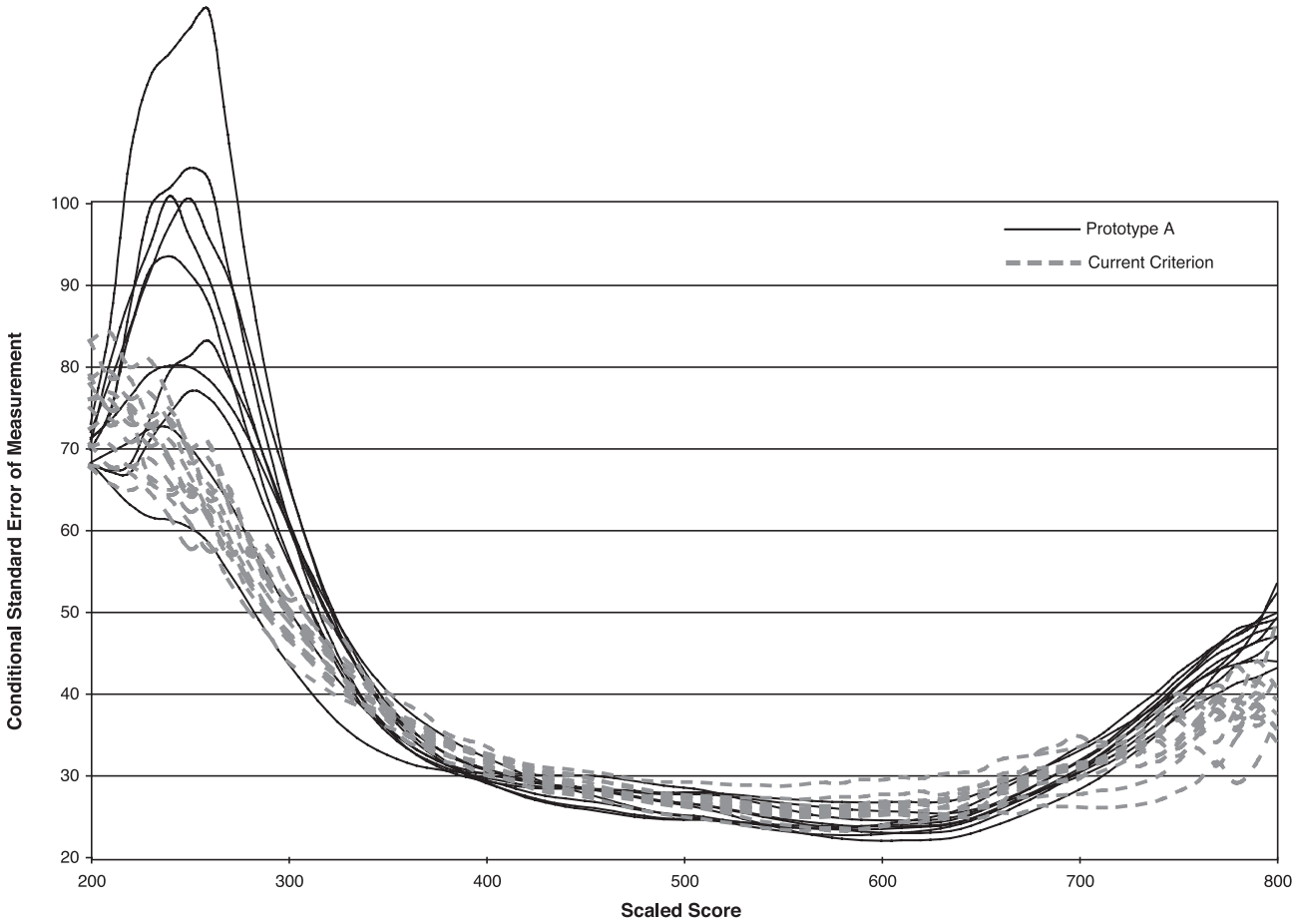

Figure 3. The scaled score CSEMs: Multiple versions of Prototype A compared to verbal sections of 13 recent SAT tests.

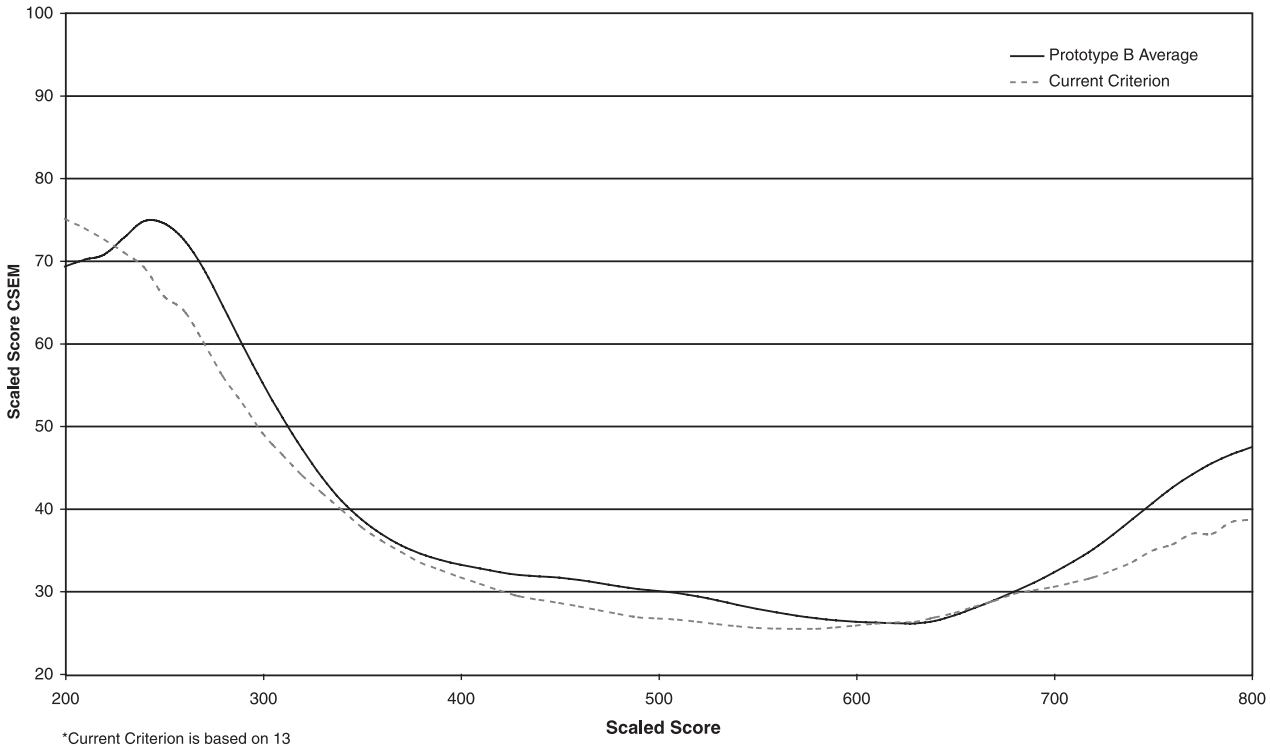

Figure 4. The average scaled score CSEMs: Prototype B compared to the SAT–V.

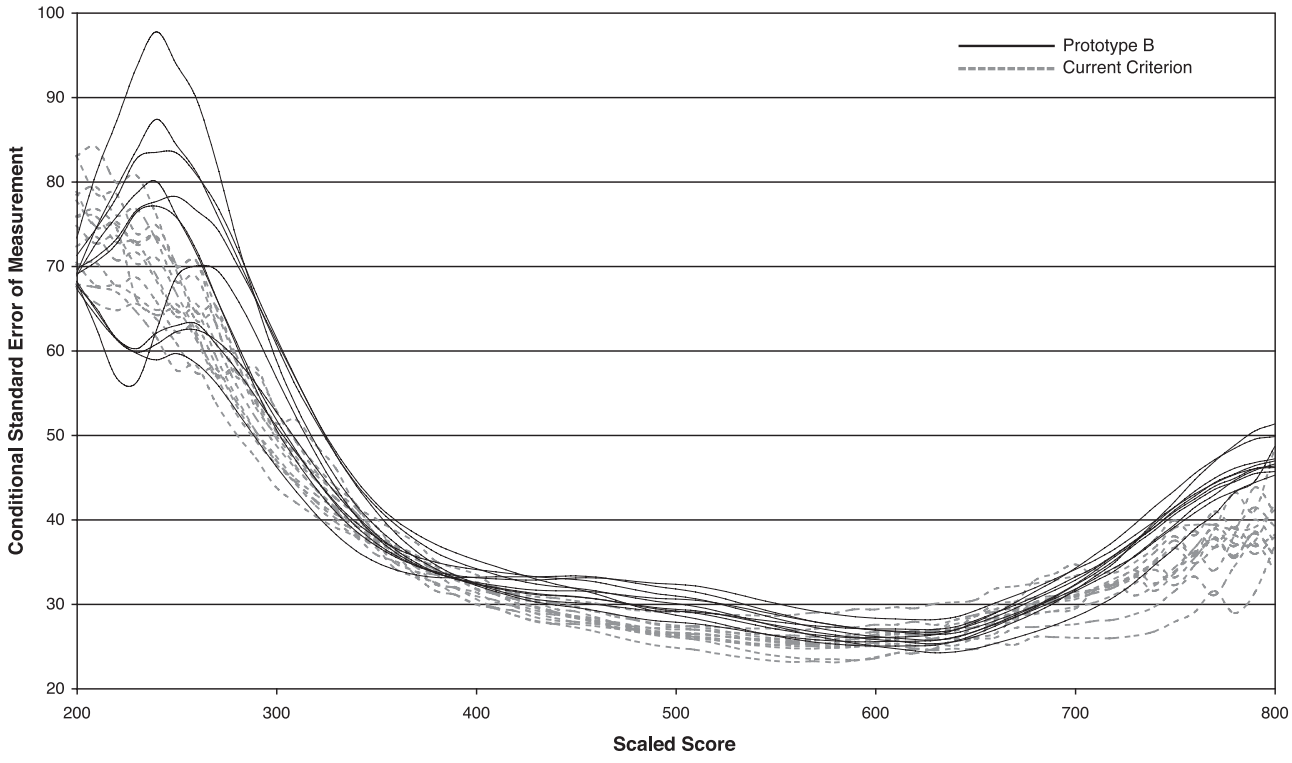

Figure 5. The scaled score CSEMs: Multiple versions of Prototype B compared to verbal sections of 13 recent SAT tests.

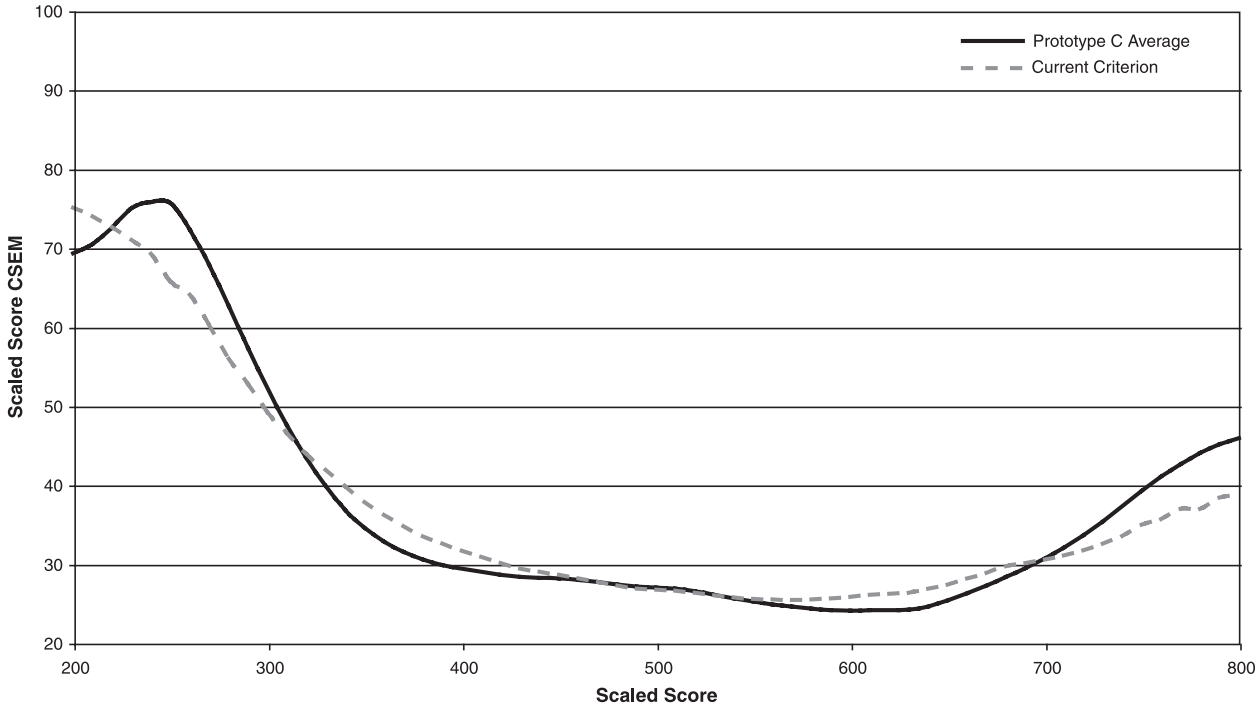

Figure 6. The average scaled score CSEMs: Prototype C compared to the SAT–V.

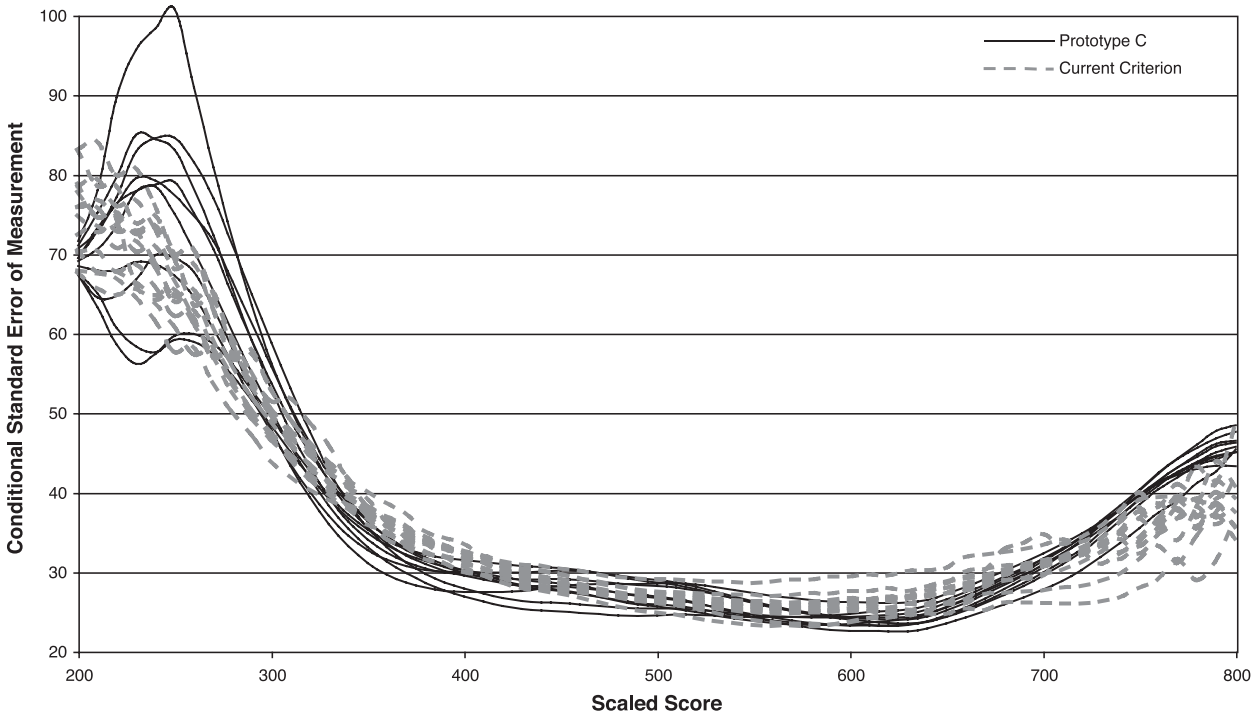

Figure 7. The scaled score CSEMs: Multiple versions of Prototype C compared to verbal sections of 13 recent SAT tests.

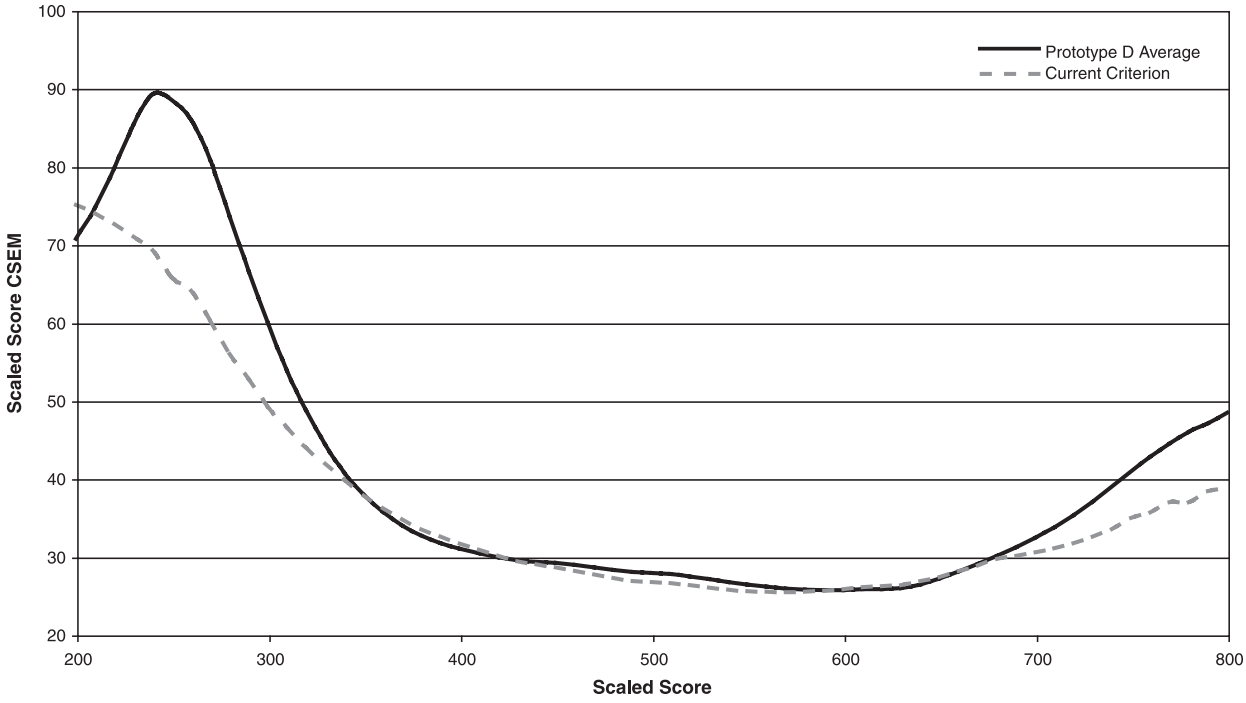

Figure 8. The average scaled score CSEMs: Prototype D compared to the SAT–V.

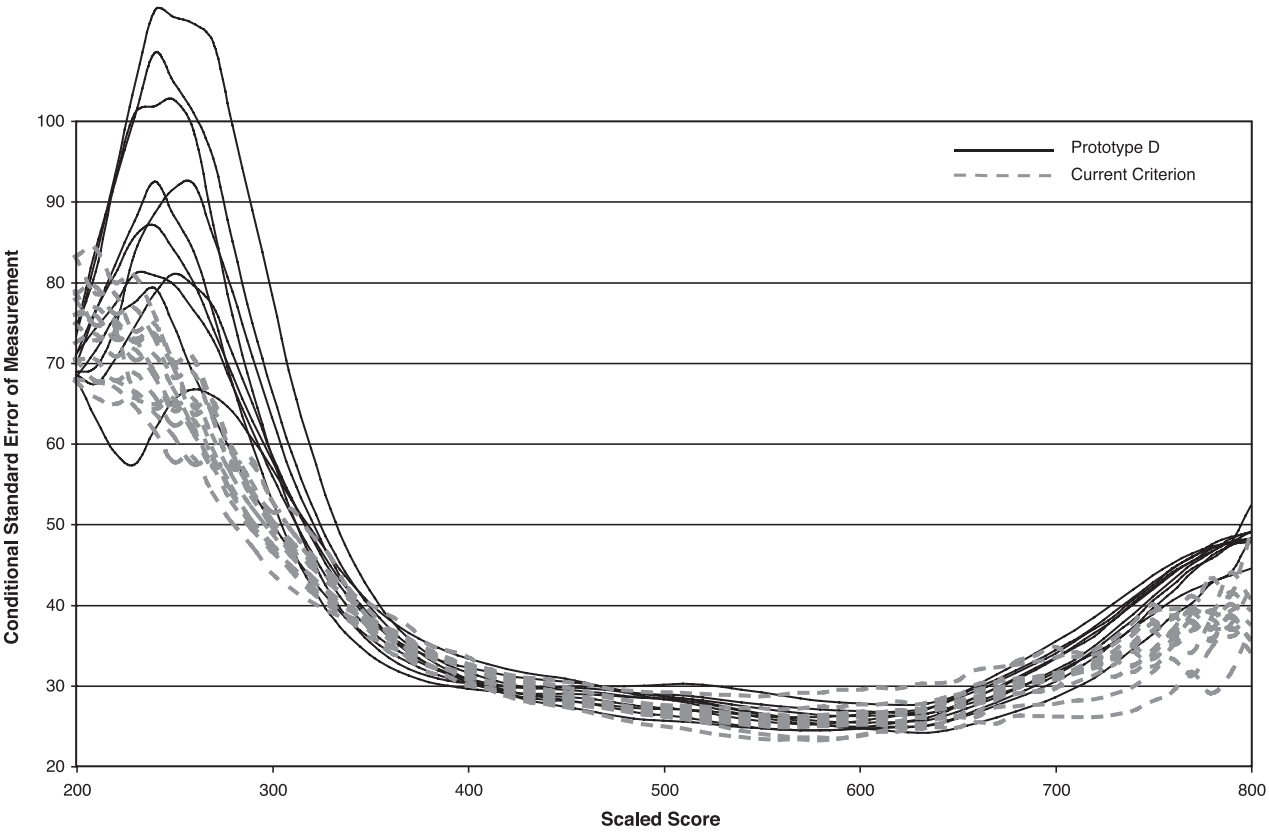

Figure 9. The scaled score CSEMs: Multiple versions of Prototype D compared to verbal sections of 13 recent SAT tests.

of the score range but showed the same trend observed for Prototype B at the extremes of the score scale.

Plots of CSEMs for Prototype D are found in Figures 8 and 9. The results of Prototype D appeared to agree closely with the criterion CSEMs throughout the mid-portion of the score range and to show the same increase in CSEMs at the extremes of the score scale as shown by other prototypes.

Reliabilities and SEMs. Table 4 shows scaled score reliability and SEM for each prototype. For all statistics found in Table 4, the average value across the multiple versions of the prototypes and the minimum and maximum values are presented for each prototype. The average reliability estimates of prototypes ranged between .91 and .93. These values can be compared to the average reliability of the SAT–V of .91, with individual values ranging between .90 and .93.

Given all other things equal, a longer test will be more reliable. The scaled score reliability estimates of all four prototypes, however, were quite high compared to the SAT–V, even for those prototypes with a considerably reduced number of items (Prototypes B and D). This may be attributed to several factors: First, the test would be more internally consistent with one item type dropped, which would result in a higher reliability. Second, compared to SC and CR items, AN items have relatively lower reliabilities. The average item type reliability estimates, based on the 13 criterion forms, are .85, .79, and .72 for CR, SC, and AN, respectively. The new critical reading section would be more reliable as a result of dropping less reliable items and retaining the more reliable items.

The data showed that the average scaled score SEM for the prototypes range between 30 and 33. These values were very close to the average SEM of the SAT–V (30). The range of SEMs for each prototype was similar to the range of SEMs for the criterion, with Prototypes B and D having slightly larger SEMs. In this case, the larger SEMs were associated with the shorter test.

TABLE 4

Reliability Estimates and Standard Error of Measurement (SEM) of Scaled Scores for the SAT–V and Prototypes

	Prototype				
	Current	*A*	*B*	*C*	*D*
Reliability (IRT)					
Mean	0.91	0.92	0.91	0.93	0.92
Minimum	0.90	0.92	0.91	0.93	0.91
Maximum	0.93	0.93	0.92	0.93	0.92
SEM					
Mean	30	31	33	30	32
Minimum	29	30	32	29	31
Maximum	32	32	34	31	34

Note: Current criteria are based on 13 SAT–V forms.

Discussion

The results of Phase 1 indicated that it is possible to construct revised test sections that are as reliable and that have overall scaled score SEMs that are similar to the SAT–V without using AN items. However, none of the prototypes, even a prototype that contained 10 analogy items, produced simulated tests with CSEMs that are as small as those produced by the SAT–V for scores below 300 or above 700.

A possible cause of this problem is that when the delta specification for each prototype was set, the number of items at each delta level was proportionally reduced based on the current specifications. Because the current delta specification is a unimodel distribution, there are fewer items over the ends of the distribution. Therefore, when the number of items at a particular level was reduced from 4 to 2, for example, it could result in very different CSEMs. This difference caused by the reduced number of items, on the other hand, would not have a significant effect on the middle of distribution where there are many more items (e.g., the number of items was reduced from 14 to 13).

Because the reduced number of items may be the cause of the larger CSEMs, it was decided to run additional simulations with more items at the ends of the delta distribution in an attempt to reduce the CSEMs in the affected regions, particularly for scaled scores above 700. Two of the most promising prototypes (A and D) were selected as the basis for additional simulations because Prototypes A and D produced smaller CSEMs throughout the mid-portion of the score range. In addition, ETS content experts were asked to evaluate each of the prototypes according to the preestablished nonpsychometric criteria described above, such as face validity, educational relevance, etc. Both Prototypes A and D received support.

Phase 2

Method

The simulations were continued during Phase 2. Prototype A and Prototype D were selected as the basis for additional simulations. Prototype A contained only a mixture of CR and SC items, and Prototype D contained CR items, SC items, and the new item type—DR items. The content configuration (mix of item types)

for these two prototypes remained the same. However, in order to increase the precision of measurement for scores above 700 and below 300, the distributions of item difficulties were altered slightly. Without changing the mean difficulty of the test, a few more difficult items, balanced with a few less difficult items, were added to the delta specifications for the two prototypes. The number of middle difficulty items was reduced to keep the total number of items unchanged.

Five simulated forms of each revised prototype were assembled for a total of 10 additional experimental test versions. As was the case for Phase 1 of the study, the mean and standard deviation of the item statistics (equated delta, *r*-biserial), and the test statistics (reliability estimates, SEMs, and CSEMs) were produced for each of the five versions of revised Prototypes A and D. The results were compared to those from Phase 1, as well as to the current criteria.

Results

Revised Delta Distribution

Table 5 contains the specified distributions of equated deltas for the SAT–V and revised Prototypes A and D. The delta distributions of the original Prototypes A and D are provided for the purpose of comparison. Compared to the delta distributions for the original prototypes, the delta distributions for revised prototypes were close to those of the original prototypes and very similar to the SAT–V. It should be recalled that an important goal of the current study is to evaluate competing prototypes for possible revisions to the SAT–V that maintain the overall difficulty level and the score reporting scale of the current tests. Consequently, the data shown in Table 5 have important implications for the study results.

TABLE 5

Specified Delta Distributions for the SAT–V, Original and Revised Prototypes

Delta Level (Specified)	*Current*	*Prototype A*		*Prototype D*	
		Original	*Revised*	*Original*	*Revised*
>=19	-	-	-	-	-
18	-	-	-	-	-
17	-	-	-	-	-
16	1	1	1	1	1
15	4	3	4	3	4
14	6	6	6	5	6
13	8	8	8	7	7
12	12	11	10	10	9
11	14	13	13	13	12
10	12	11	10	10	9
9	9	8	8	8	7
8	7	6	6	6	6
7	4	4	4	3	4
6	1	1	2	1	2
<=5.9	-	-	-	-	-
Number of Items	78	72	72	67	67
Mean	11.4	11.4	11.4	11.4	11.4
S.D.	2.2	2.2	2.3	2.2	2.4

TABLE 6

Summary of Item Statistics for the SAT–V and Revised Prototypes

	Current	*Revised Prototype*	
		A	*D*
Number of Items	78	72	67
Equated Delta			
Mean	11.4	11.4	11.4
S.D.	2.2	2.2	2.4
r-biserial			
Mean	0.51	0.53	0.52
S.D.	0.10	0.10	0.10

Note: Current criteria are based on 13 recent SAT–V forms.

Item Statistics

Table 6 provides information about the item statistics for the revised prototypes. The mean and standard deviation of equated deltas and the mean and standard deviation of *r*-biserial obtained for the revised prototypes were very similar to those for the criterion.

Test Statistics

Plots of IRT scaled score CSEM. Plots of IRT scaled score CSEMs for the revised Prototypes A and D can be found in Figures 10 and 11. These plots show across the scaled score range, the average CSEM values of each revised prototype compared to the average CSEM values of the criterion, and compared to the average CSEM values of the original prototypes.

An examination of the average CSEM for revised Prototype A compared to the current criterion and the original Prototype A (Figure 10) shows that it is still the case that the revised prototype had larger CSEM than the current criterion for scaled scores below about 300 and above approximately 700. However, the discrepancy between the revised prototype and criterion CSEMs appeared to be somewhat reduced compared to the discrepancy between the original prototype and the criterion CSEM.

Figure 11 shows plots of CSEMs for revised Prototype D. It is still the case that the average of revised Prototype D appeared to result in higher CSEMs than the criterion in the ends of the score range. However, when compared to the CSEMs for the original Prototype D, it can be seen that revising the equated delta specifications has resulted in some improvement in

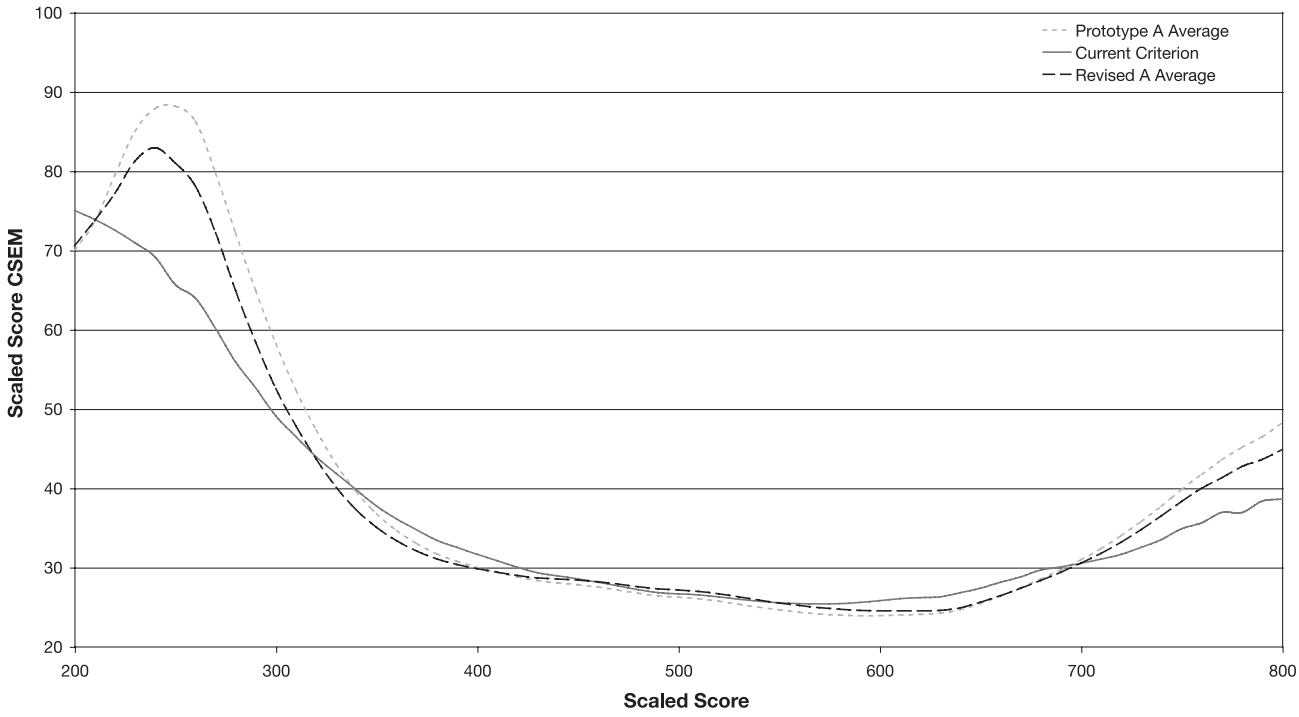

Figure 10. The average scaled score CSEMs: Revised Prototype A compared to the SAT–V and the original Prototype A.

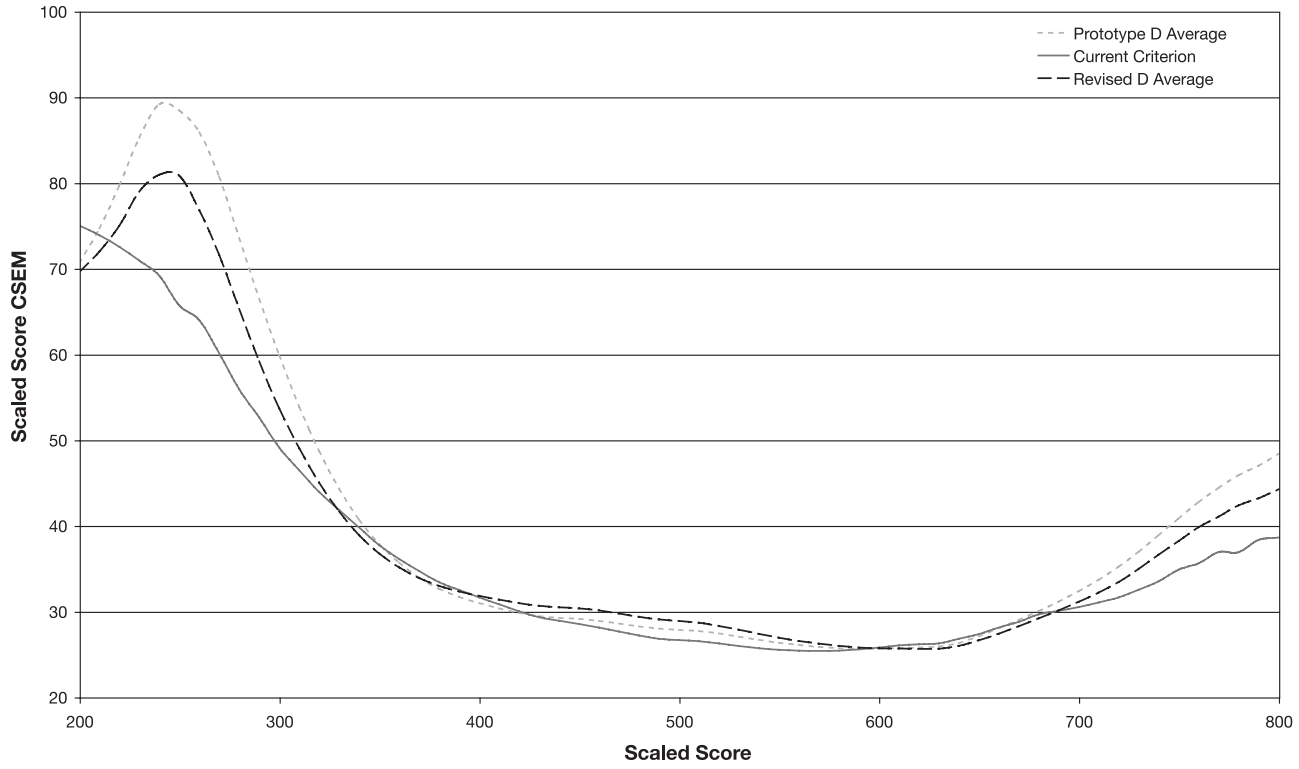

Figure 11. The average scaled score CSEMs: Revised Prototype D compared to the SAT–V and the original Prototype D.

TABLE 7

Reliability Estimates and Standard Error of Measurement (SEM) of Scaled Scores for the SAT–V and Revised Prototypes

	Current	Revised Prototype	
		A	D
Reliability (IRT)			
Mean	0.91	0.93	0.92
Minimum	0.90	0.92	0.91
Maximum	0.93	0.93	0.92
SEM			
Mean	30	30	32
Minimum	29	29	31
Maximum	32	31	33

Note: Current criteria are based on 13 SAT–V forms.

the measurement power of the test for scores in the upper and the lower portions of the score scale.

SEM and reliability. Table 7 shows test statistics (IRT scaled score reliability and SEM) for the revised prototypes. The average value of the statistics across the multiple versions of the prototypes and the minimum and maximum values were presented for each prototype.

It can be seen that the average reliability estimates ranged between .92 and .93 under revised Prototype A, and ranged from .91 to .92 under revised Prototype D. These values can be compared to the reliability of the criteria that ranges from .90 to .93. The scaled score reliability estimates of the revised prototypes were quite high and did not seem to be adversely impacted by the small revisions to the equated delta specifications.

The scaled score SEMs ranged from 29 to 31, and from 31 to 33 for revised Prototype A and D, respectively. These values can be compared to the scaled score SEMs of the SAT–V that ranges between 29 and 32.

Discussion

The results of the analyses indicated that there was a small reduction in the size of the CSEM for the revised prototypes for scaled scores above 700 and for scaled scores below 300. The overall SEM and the test reliability for the five versions of each revised prototype were similar to those of the SAT–V and of the 10 versions of each original prototype.

The results of Phase 2 supported the hypothesis that the reduced number of easy and hard items caused increased CSEMs in the ends of the score scale. Another possible cause for the increased CSEMs is that different item types play different roles in causing measurement errors. Assuming that we have a "verbal" test section containing only AN items, only SC items, or only CR items, what CSEM curves will we obtain? Phase 3 was designed to explore CSEM differences caused by different item types. Three hypothetical tests were constructed: An all-AN test, an all-SC test, and an all-CR test. The CSEMs for each of the hypothetical tests were analyzed and compared.

Phase 3

Method

Estimation of the Length and Delta Distribution of Three Hypothetical Tests

ETS content experts estimated the number of items of each item-type test that is 75 minutes long by taking into account the issues such as fatigue and speededness of each item type. Therefore, each test was adjusted to a comparable test length. The estimated length of each item-type test was: 60 AN items, 62 CR items, and 50 SC items.

The total item pool was divided into three distinct subsets: AN, SC, and CR subpool. Table 8 shows the delta distributions in each subpool. The percentage columns indicated the percentage at each delta level, based on the corresponding total number of items. As

TABLE 8

Specified Delta Distributions for Different Item Types in the Item Pool

Delta	*AN*	*% of pool*	*CR*	*% of pool*	*SC*	*% of pool*
>=19	2	0.4%	3	0.4%	2	0.4%
18	0	0.0%	1	0.1%	0	0.0%
17	0	0.0%	1	0.1%	0	0.0%
16	17	3.4%	4	0.6%	8	1.8%
15	49	9.9%	21	3.0%	17	3.8%
14	40	8.1%	50	7.2%	36	7.9%
13	62	12.6%	86	12.3%	42	9.3%
12	61	12.3%	144	20.7%	66	14.6%
11	58	11.7%	147	21.1%	68	15.0%
10	57	11.5%	108	15.5%	76	16.8%
9	44	8.9%	83	11.9%	60	13.2%
8	52	10.5%	37	5.3%	48	10.6%
7	34	6.9%	9	1.3%	24	5.3%
6	16	3.2%	3	0.4%	6	1.3%
<=5.9	2	0.4%	0	0.0%	0	0.0%
Total	494		697		453	
Mean	11.9		12.0		11.6	
S.D.	6.2		6.1		6.3	

Note: The percentage is based on the total number of items for each item type.

TABLE 9

Estimation of Hypothetical Item-Type Test

Delta	*AN*	*% of test*	*CR*	*% of test*	*SC*	*% of test*
>=19	0	0.0%	0	0.0%	0	0.0%
18	0	0.0%	0	0.0%	0	0.0%
17	0	0.0%	0	0.0%	0	0.0%
16	2	3.3%	0	0.0%	1	2.0%
15	6	10.0%	2	3.2%	2	4.0%
14	5	8.3%	5	8.1%	4	8.0%
13	8	13.3%	8	12.9%	5	10.0%
12	8	13.3%	13	21.0%	7	14.0%
11	7	11.7%	13	21.0%	7	14.0%
10	7	11.7%	10	16.1%	8	16.0%
9	5	8.3%	7	11.3%	7	14.0%
8	6	10.0%	3	4.8%	5	10.0%
7	4	6.7%	1	1.6%	3	6.0%
6	2	3.3%	0	0.0%	1	2.0%
<=5.9	0	0.0%	0	0.0%	0	0.0%
Total	60		62		50	
Mean	11.7		11.7		11.3	
S.D.	2.7		1.8		2.3	

can be seen, the delta means and standard deviations across three item types were quite close to each other. However, there were more AN items over the ends of the scale than CR or SC items. For example, at delta level 16, the number of each type of items were: 17 AN (3.4 percent), 4 CR (0.6 percent), and 8 SC (1.8 percent); and at delta level of 15, there were 49 AN (9.9 percent), 21 CR (3.0 percent), and 17 SC (3.8 percent). A similar pattern can be observed at the lower end of the delta scale.

Table 9 shows the delta distribution for each item-type test. The delta distributions were obtained by proportionally reducing the number of items at each delta level to reflect the reduced total number of items in each item-type test.

Computation of CSEM for Three Hypothetical Tests

Item parameter estimates for all items in the three subpools were placed on the same base form scale to which all individual simulation forms were scaled in order to get CSEMs for each subpool. The CSEMs were then computed for each hypothetical test by using the following formula:

$$(6)\quad CSEM_{Test} = CSEM_{Pool}\sqrt{\frac{N_{Test}}{N_{Pool}}},$$

where N_{Test} is the total item number in each typical item-type test (e.g., 60 for an AN test), and N_{Pool} is the total item number in each subpool (e.g., 494 for the AN item pool).

Formula (6) was derived from Formula (2),

$$CSEM\left(RS_j\middle|\theta_j\right) = \sqrt{\sum_{i=1}^{K_T} P_i\left(\theta_j\right)Q_i\left(\theta_j\right)},$$

where K_T is the total number of items on the test. Formula (2) makes clear that (under the binomial model) the squared CSEM is an additive function of the conditional variance of each item. Given the total pool of items, we would compute the squared CSEM for score level *j* by summing the conditional variances for all items in the pool:

$$(7)\quad CSEM^2_{Pool}\left(RS_j\middle|\theta_j\right) = \sum_{i=1}^{N_{Pool}} P_i\left(\theta_j\right)Q_i\left(\theta_j\right).$$

Given the squared CSEM for the total pool of items, we could estimate the conditional variance for a single item by dividing the squared CSEM by the number of items in the pool,

$$(8)\quad CSEM^2_{Item} = \frac{CSEM^2_{Pool}}{N_{Pool}}.$$

Although not explicitly stated in the formula (for ease of reading) the CSEM is understood to be for score level *j*, conditioned on Θ. From this value, we could estimate the squared CSEM for any length test by multiplying the conditional item variance by the number of test items,

$$(9)\quad CSEM^2_{Test} = N_{Test}\cdot CSEM^2_{Item} = CSEM^2_{Pool}\cdot\frac{N_{Test}}{N_{Pool}}.$$

The use of this formula rests on the assumption that all tests are composed of items that are representative of the total pool (in terms of difficulty). Note, too, that this derivation would still be valid if Formula (3) had been used instead of Formula (2).

Results

The results of these analyses are presented in Figure 12. These plots exhibit the relative CSEMs across the 200 to 800 SAT scale for these three hypothetical tests. It is clear that the AN test provided the most uniform measurement over the entire score range, and yielded the smallest CSEMs for scaled scores in the upper range (above 700) and the lower range (below 300) of the score distribution. In contrast, both the CR and SC tests measured the middle of the score range (400–600) best, and tended to provide poorer measurement over the ends.

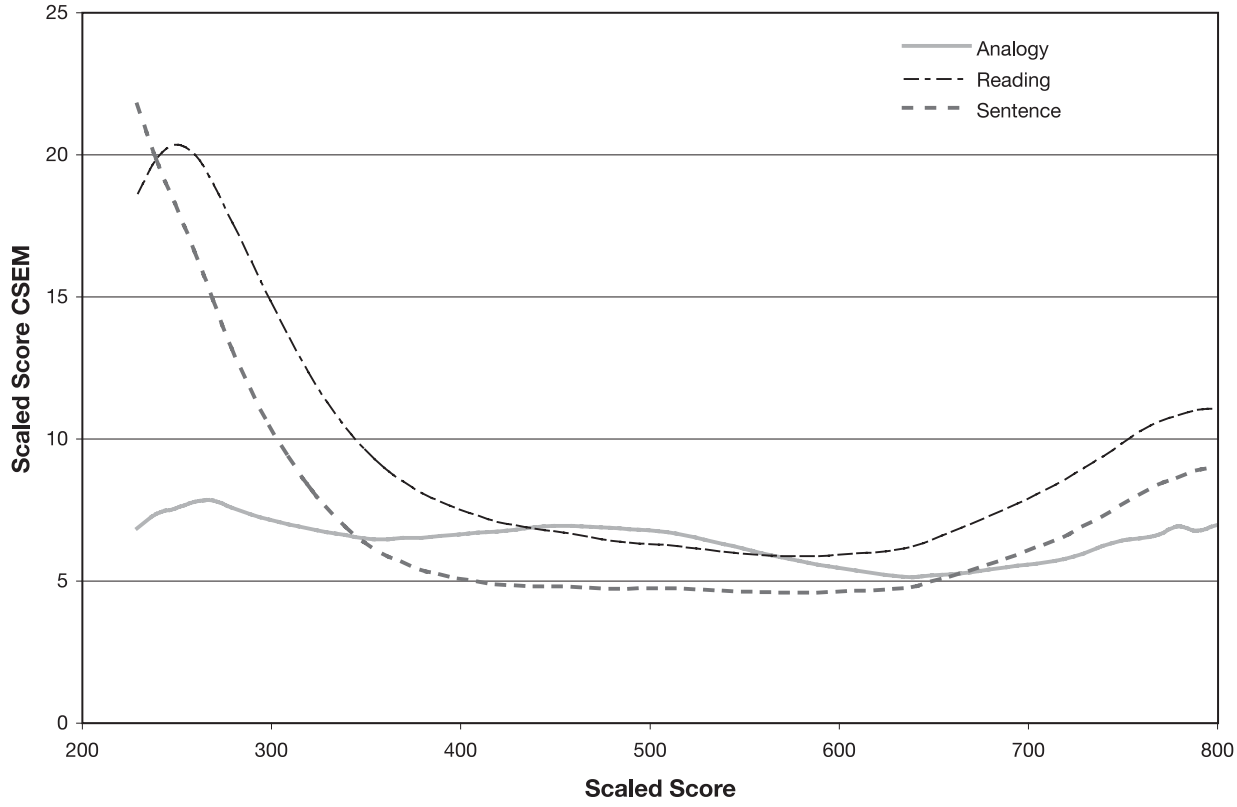

Figure 12. The average scaled score CSEMs for tests composed solely of a single SAT–V item type.

These results further illuminated the results from Phases 1 and 2: the prototypes without AN items resulted in higher CSEMs in the ends of the score range. A possible reason is that most of the very difficult or very easy items are analogy items. As examined above, there were more AN items over the ends of the delta scale range in the item pool. Similarly, when the number of items were proportionally reduced at each delta level in each item type test, there were more difficult and easy AN items over the ends of the scale. As shown in Table 9, for example, there were 2 (3.3 percent) AN items at delta 16, where there were no CR items and there was 1 (2.0 percent) SC item. The differences were exaggerated at delta level of 15, where there were 6 AN items (10.0 percent), and only 2 CR (3.2 percent) and 2 SC items (4.0 percent) each. The lower end of the scale demonstrated the same pattern: from delta 8 through delta 6 (or below 5.9), there were 12 AN items, 4 CR items, and 9 SC items.

The results indicated that if the new SAT critical reading section is to be constructed without AN items, extra care will need to be taken to ensure that the item pool contains sufficiently difficult and discriminating CR and SC items.

General Discussion

The purpose of this study was to explore the possibility of configuring the new SAT critical reading section without AN items. Four prototypes were designed and 10 versions of each prototype were constructed and analyzed.

This study employed an application of item-response theory to evaluate the results of reconfiguring a test section, such as the SAT–V, prior to actually field-testing the proposed revisions. The study was possible because data (item responses) existed for all items comprising the revised forms. After forming a pool of items with item parameter estimates on the same scale, it was possible to simulate the effects of various proposed configurations of the SAT–V on test statistics such as reliability and conditional standard errors of measurement.

The results of the initial analyses indicate that it is possible to construct prototypes that are as reliable and that have overall scaled score SEM that are similar to the SAT–V without using AN items. However, none of the prototypes, even Prototype C that contained 10 analogy items, produced test simulations with CSEMs

that are as small as those produced by the SAT–V for scores below 300 or above 700.

The prototypes built without analogy items that appear to have the most promise are Prototypes A and D. When asked to evaluate the prototypes according to the nonpsychometric criteria used for the study, ETS content experts felt both of these prototypes were acceptable. Since Prototypes A and D were favored by ETS content staff, further simulations were carried out for these prototypes using slightly revised equated delta specifications. Five additional versions of each of these two prototypes were simulated. The revised specifications (a few high delta items were added) did appear to reduce the size of the CSEMs for scores above 700 for both prototypes, even though it is still the case that both revised prototypes have larger CSEMs than the current criteria in the ends of the scaled score range.

Exploratory analyses that focused on the relative size of the CSEMs produced by an "all-AN item," "all-CR item," or "all-SC item" test showed that a test constructed solely from analogies would result in the smallest CSEMs for scores at the upper end and lower end of the scale range. These results indicate that special care will need to be taken to ensure that a sufficient number of highly difficult and highly discriminating CR and SC items are developed to support the needs of the new SAT critical reading section that will not include analogy items.

This simulation study provided information on psychometric characteristics and possible configurations of the prototypes before carrying out an expensive field trial. On the other hand, some information such as test speededness cannot be obtained from the simulations. It is recommended that the draft prototypes designed in this study undergo further investigation by collecting and analyzing data from intact prototypes given at actual administrations to real examinees. The two revised prototypes, A and D, should be tried out with real test-takers and the results of these analyses evaluated prior to a final decision regarding revisions to the current SAT verbal section. In addition, it is recommended that the feasibility of building and maintaining a verbal pool with a sufficient number of difficult and discriminating reading passage and sentence completion items be considered if either of the verbal prototypes is adopted.

References

Bridgeman, B., Cahalan, C., & Cline, F. (May 2003). *Time requirements for different item types proposed for use in the revised SAT.* Draft Report, Princeton, NJ: Educational Testing Service.

College Board (2002). *The new SAT: Implemented for the class of '06*. PowerPoint slides posted on http://www.collegeboard.com/prod_downloads/about/newsat/newsat_presentation.ppt.

Cook, L.L. & Petersen, N.S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244.

Dorans, N. (1984). *Approximate IRT formula score and scaled score standard error of measurement at different ability levels*. Statistical Report, SR-84-118, Princeton, NJ: Educational Testing Service.

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

Stocking, M.S. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Stocking, M.L. & Swanson, L. (1992). *A method for severely constrained item selection in adaptive testing*. Research Report, RR-92-37, Princeton, NJ: Educational Testing Service.