DOCUMENT RESUME

ED 397 075 TM 025 070

AUTHOR	Lawrence, Ida M.
TITLE	Estimating Reliability for Tests Composed of Item Sets.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	College Entrance Examination Board, New York, N.Y.
REPORT NO	ETS-RR-95-18
PUB DATE	Jul 95
NOTE	29p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); High Schools; *Multiple Choice Tests; *Reading Tests; *Reliability; *Test Items; Verbal Tests
IDENTIFIERS	*Item Dependence; Scholastic Aptitude Test; *Scholastic Assessment Tests

ABSTRACT

This study examined to what extent, if any, estimates of reliability for a multiple choice test are affected by the presence of large item sets where each set shares common reading material. The purpose of this research was to assess the effect of local item dependence on estimates of reliability for verbal portions of seven forms of the old and seven forms of the new Scholastic Aptitude Test, where the new test contains larger item sets associated with reading passages. Estimates based on a single administration of the test (estimates based on internal consistency and estimates based on covariances among parts) were compared to estimates based on two administrations of the test. When adjusted for a fixed standard deviation, estimates based on covariances among parts tended to be similar to estimates based on parallel forms. Both types of estimates were lower than the internal consistency estimates. (Contains 3 figures, 5 tables, and 12 references.) (Author/SLD)

RR-95-18

RESEARCH REPORT

ESTIMATING RELIABILITY FOR TESTS COMPOSED OF ITEM SETS

Ida M. Lawrence

Educational Testing Service
Princeton, New Jersey
July 1995

Estimating Reliability for Tests Composed of Item Sets

Ida M. Lawrence[1]

Educational Testing Service

[1]The author thanks Miriam Feigenbaum, Nancy Feryok, and Robert Kelly for their help producing and analyzing the data, and Dan Eignor, Charlie Lewis, Skip Livingston, Gary Marco and Nancy Wright for their reviews of earlier versions of this paper. Support for this research was provided by the College Board.

Abstract

To what extent, if any, are estimates of reliability for a multiple choice test affected by the presence of large item sets where each set shares common reading material? The purpose of this research was to assess the effect of local item dependence on estimates of reliability for verbal portions of the old and new SAT, where the new test contains larger item sets associated with reading passages. Estimates based on a single administration of the test (estimates based on internal consistency and estimates based on covariances among parts) were compared to estimates based on two administrations of the test. When adjusted for a fixed standard deviation, estimates based on covariances among parts tended to be similar to estimates based on parallel forms. Both types of estimates were lower than the internal consistency estimates.

Estimating Reliability for Tests Composed of Item Sets

To what extent, if any, are estimates of reliability for a multiple choice test affected by the presence of large item sets where each set shares a common reading passage? The research described in this paper sought to shed light on this question in an applied setting.

This research was prompted by concerns raised by Wainer and Thissen (in press) and, Sireci et al (1991), who found that estimates of reliability based on internal consistency approaches may be misleading if the test is composed of reading passages with large sets of items. When more rather than fewer items are related to a single passage, the dependence among items is increased, consequently, internal consistency estimates of reliability may be inflated relative to estimates of reliability based on correlations between alternate forms of the test.

According to Yen (1993), if several items are attached to the same reading passage or common stimulus material, LID (Local Item Dependence) can occur. The dependence among items associated with a passage or common stimulus may be a consequence of a test taker's particular interest in a topic, or the result of background or additional knowledge about the content of the passage or stimulus.

This research used data from the verbal portions of the old and new SAT. The old SAT verbal test contained four item types (passage-based reading questions, sentence completion questions, analogy questions, and antonym questions). Sections in the old test contain the following items:

Section 1 -- 10 sentence completion questions, 10 analogy questions, 15 antonym questions, and 10 reading questions (based on 2 passages).

Section 2 -- 5 sentence completion questions, 10 analogy questions, 10 antonym questions, and 15 reading questions (based on 4 passages).

The new verbal test contains three item types (passage-based reading questions, sentence completion questions, and analogy questions). Sections in the new test contain the following items:

Section 1 -- 10 sentence completion questions, 13 analogy questions, and 12 or 13 reading questions (based on 1 passage).

Section 2 -- 9 sentence completion questions, 6 analogy questions, and 15 or 16 reading questions (based on 2 passages).

Section 3 -- 11, 12, or 13 reading questions (based on 1 passage).

The new test differs from the old test in that it places greater emphasis on reading. Approximately 50% of the questions in the new test are passage-based, compared with 26% in the old test. The verbal portion of the old SAT limited the number of questions following a passage to 5 or 6 questions. The verbal portion of the new SAT contains reading passages followed by as many as 12 or 13 questions. Figure 1 shows the breakdown of item types and numbers of questions associated with passages in the old test and the new test.

The old and new test also differ with respect to test length and the specified distributions of item difficulty, differences that may affect reliability. Figure 2 shows the specified distributions of item difficulty (in terms of equated deltas[1]) for the old test (85 items) and the new test (78 items). Specifications for the old test called for a fairly bimodal distribution of item difficulty. In contrast, the statistical specifications for the new test require a more unimodal distribution of item difficulty, with a larger proportion of

[1]The delta index is based on the percent of test takers answering the item correctly (i.e., p-value), where 1 minus the p-value is converted to a normalized z-score and then transformed to a scale with a mean of 13 and a standard deviation of 4. Raw delta values are converted to equated delta values to estimate item difficulty for a reference population.

middle difficulty items and fewer very easy and very hard items. The average overall difficulty of items in the old test and the new test is the same. (See Lawrence & Schmitt, 1994, for details about the process of setting statistical specifications for the new SAT.)

This research sought to answer the following questions:

1. Is the new SAT as reliable as the old SAT?
2. What are the content and statistical differences between the new and old SAT, and how could the differences be expected to affect reliability?
3. What are appropriate methods for estimating reliability, what would cause each method to underestimate or overestimate reliability, and would these factors apply equally to the old and new SAT?

Data Source

Data for this research came from two sources:

National Test Administrations

Reliability estimates were obtained for seven forms of the old SAT and seven forms of the new SAT. Analyses were based on representative samples of test takers (high school juniors and seniors) from national administrations. Sample sizes and summary statistics for reliability estimates based on national test administrations are presented in Table 1.

Test Takers who Took Parallel Forms at National Administrations

Data came from test takers who took the old SAT in March 1993, May 1993, or June 1993 and repeated the test at one of these three 1993 administrations, or took the test in March 1992, May 1992, or June 1992 and repeated the test at one of these three 1992 administrations. With no more than two months transpiring between test administrations, these data were used to estimate parallel-form reliability of the old test. Similar data from the March 1994, May 1994 and June 1994 administrations was used to estimate the parallel-form reliability of the new test. Sample sizes and summary statistics for reliability estimates based on national test administrations are presented in Table 2. The last two columns show differences between means and ratios of standard deviations for the first and second test taken.

Estimates of Reliability

When assessing test reliability, several approaches are possible. Ideally, estimates of reliability are derived from scores on parallel forms of a test. With this approach, referred to in this research as the parallel-form approach, the estimate of reliability is the correlation between parallel forms of a test taken one or two months apart. Another, less ideal, approach is to obtain an estimate of reliability from a single administration of the test. Although estimates from a single test administration apply only to the form being analyzed, the intent is to approximate reliability estimates based on more than one test administration.

Measures of reliability based on a single test administration do not take into account lack of parallelism among forms. A correlation between parallel forms that is considerably lower than the estimates of reliability based on a single test administration may indicate the presence of measurement error that is due to differences in content sampling between forms of the test. In addition, reliability estimates based on a single test administration do not take into account differences in testing conditions across administrations of the test, or test taker factors such as whether the test taker was ill, and so on. In addition, reliability estimates based on a single test administration do not take into account the effects of score equating error.

The formulas used to estimate reliability from a single test administration are presented in Figure 3. The four indices used in this research are described below.

Dressel KR-20 (Alpha)

Estimates based on an internal consistency method of estimating reliability, such as the Dressel (1940) adaptation[2] of the Kuder Richardson-20 (KR-20) estimate (equivalent to coefficient alpha), indicate the consistency of performance of test takers on items within a test. The value of Dressel-KR-20 depends on the average inter-item correlation and the number of items in the test. Relative to estimates based on parallel forms, this index is inflated when the test is speeded and deflated when the test measures more than a single underlying dimension. It also may be inflated when items depend on common stimuli (e.g., reading passages). Dressel KR-20 was computed for each separately timed section in the old and new test.

Variance Components

The variance components measure was used to estimate reliability for the total test from the Dressel KR-20 raw score standard errors of measurement for each separately timed section of the test. It is assumed that the sections are parallel in content but not necessarily in difficulty or test length. Test speededness and multidimensionality within section have the same effect on the variance components estimate as on the Dressel KR-20 estimate.

Angoff-Feldt and Kristof

In addition to estimating reliability from the variance components of the separately timed sections, total test reliability can be estimated by the Angoff-Feldt procedure (Angoff, 1953; Feldt, 1975) when the test is composed

[2]The Dressel (1940) adaptation of KR-20 is for formula scored tests.

of two separately timed sections. A slightly different formula, developed by Kristof (1974), may be used to estimate total test reliability for tests composed of three separately timed sections. The Angoff-Feldt and Kristof procedures assume that the separately timed sections are parallel in content but not necessarily in difficulty or test length. These estimates are more accurate than a total test internal consistency measure (e.g., the variance components index) when the test is speeded or measures more than one dimension. Both of these approaches, however, will underestimate reliability when the separately timed sections are not strictly congeneric. Whether each section measured the same construct was an issue for the new test, because one of the sections contains only one of the three item types (the third section contains only passage-based reading questions). In contrast, all four item types in the old SAT are represented in both of the separately timed sections.

In order to estimate reliability assuming congeneric parts, the items in new test were split into two content-homogeneous parts by treating section 2 as one part and section 1 and section 3 combined as the other part. Angoff-Feldt estimates based on the congeneric parts were compared to the Kristof estimate, which assumes that the three separately timed sections in the new test are congeneric.

As a result of using items in separately timed sections of the new test to form congeneric parts for the Angoff-Feldt reliability estimate, items associated with a given passage remained intact. Feldt and Brennan (1989) point out that there are pros and cons associated with intact item sets. To point out the issues, Feldt and Brennan provide an example of a reading test where different types of passages are included in each edition of the test, and parallel forms are matched in terms of this passage typology. With this kind of configuration, splits based on intact passages will tend to be more different than the differences among passages in parallel forms and the associated reliability estimate. However, separating items within passages will tend to increase the correlation among parts, thus inflating the reliability estimate.

Both the old and new test contain reading passages, from different content areas (e.g., Humanities, Natural Sciences, Social Sciences, Narration). Since there are few or no within-test replications for the content areas associated with the passages, the Angoff-Feldt or Kristof methods could result in reliability estimates that are lower than that which would be obtained with a parallel-forms approach. On the other hand, the content areas are fairly broad, so there is no expectation that forms would actually be equalized with respect to passage content.

Relative to estimates based on parallel forms, the Angoff-Feldt and Kristof estimates do not account for the variance component due to growth (see discussion of growth below). The actual size of this component is an empirical question not addressed in this research.

Analyses

Reliability Coefficients from a Single Test

Estimates of reliability based on a single test administration were obtained for the old test and new test using data from national administrations.

Reliability Coefficients from Parallel Forms

Reliability based on the parallel-forms approach was also estimated for the old test and for the new test by computing correlations between scaled (200-to-800 College Board scale) scores on a first and second testing ("repeaters"). The time interval between the first test and the second test was either one month or two months.

12

Strictly speaking, the analysis of repeater data violates an assumption of the parallel-forms approach to estimating reliability because of the possibility of growth between the first and second testing. This effect is seen in Table 2, which shows summary statistics for the first and second tests. Indeed, for the old test, scores on the second test are 6 to 21 points higher (at the mean, on the 200-to-800 College Board scale) than scores on the first test. For the new test, scores on the second test are 2 to 11 points higher than scores on the first test. The effect of practice on the reliability coefficient is not clear cut; one possibility is that practice might reduce error variance on the second test. Although there appears to be slight gain scores between the first and second test, the variability of scores on the first test and second test is fairly similar, as indicated by ratios of the standard deviation on the second test to the standard deviation of the first test; note that the ratios are close to 1.00.

Another limitation to this repeater data is that the parallel-forms estimates come from samples of self-selected test-takers who chose to repeat the test within one or two months of the first test. Thus, the samples are not representative of test takers from national administrations. In particular, the repeater samples are less variable. Standard deviations for the national administration samples range between 102 and 111 (Table 1). As can be seen in Table 2, standard deviations for the repeater samples range between 91 and 110. The likely effect of the restricted range in scores is to attenuate the parallel-form reliability coefficients.

Reliability estimates based on a single test administration are computed using data from representative samples of test takers in national administrations. In contrast, reliability estimates based on two test administrations are computed using data from self-selected samples of test takers who chose to repeat an SAT at a second administration. A comparison of the summary statistics in Table 1 and Table 2 reveals that the representative samples and the "repeater" samples are quite different. To obtain reliability estimates for equivalent samples, the coefficients based on single test administrations and two administrations were adjusted via the following formula:

(1) $$\text{reliability (adjusted)} = 1 - (\text{SEM}^2 / \text{assumed variance})$$

A standard deviation of 110 was assumed to be a constant across samples.

Results and Discussion

Estimates Based on a Single Test Administration

Reliability estimates and scaled score SEMs based on a single administration of seven forms of the old test and seven forms of the new test are shown in Table 3. Within each form, the estimates based on covariances among congeneric parts (Angoff-Feldt and Kristof) are slightly lower than the estimates based on internal consistency (variance components).

Although the new test has fewer items than the old test, the variance components reliability estimates and standard errors of measurement are similar across the old and new tests. An important factor affecting this result is that the new test is more peaked with respect to item difficulty than the old test.

The Angoff-Feldt reliability estimates for the old test are similar to the Kristof estimates for the new test, but the standard errors of measurement for the new test tend to be larger, indicating that the new test is slightly less reliable than the old test.

The variance components measures may be overestimates due to slight speededness of the individual test sections and increased item dependence due to item sets. The Angoff-Feldt and Kristof measures may be underestimates due to lack of parallelism across separately timed sections.

The close agreement between the Angoff-Feldt and Kristof measures for the new test indicates that the assumption of congeneric parts has been satisfied for the Kristof estimate. In other words, the three sections in the new test are sufficiently parallel in content to not affect the Kristof estimate relative to the Angoff-Feldt estimate, where item types are set up to be similar across parts.

Estimates Based on Two Test Administrations

Reliability estimates and SEMs based on the parallel-forms approach for the old test and new test are presented in Table 4 (coefficients for the old test and the new test are rank ordered from highest to lowest). Estimates of reliability based on a single administration of the test are slightly larger than estimates of reliability based on two administrations. Factors that are likely to be responsible for attenuating the correlation between forms are (a) scores for the repeater samples are restricted in range compared to scores for representative samples, and (b) the presence of slight score gain between the first and second test. Note that gains on the old test tend to be larger than gains on the new test.

On average, parallel-forms reliability coefficients for the old test are .01 larger than the corresponding reliability coefficients for the new test. Out of 15 coefficients for the new test, 6 are smaller than the smallest coefficient for the old test. Out of 8 coefficients for the old test, 2 are larger than the largest coefficient for the new test.

Adjusted Reliability Estimates

Table 5 shows reliability coefficients for the old and new test that have been adjusted for a fixed standard deviation of 110 (using Formula 1). The table presents estimates based on a single administration of the test (variance components, and Angoff-Feldt for the old test and Kristof for the new test) along with adjusted average parallel-form estimates (see below). Due to the nature of the repeater data, several adjusted parallel-form coefficients are available for each form (i.e., Form O1 was paired with Form O2 in one sample and with Form O3 in another sample). As a summary index for a particular form, the relevant adjusted coefficients were averaged. The adjusted variance components estimates tend to be similar for the old test and the new test. In contrast, the other adjusted reliability estimates tend to be slightly smaller for the new test than for the old test. In particular, the average adjusted parallel-form estimate is .005 lower for the new test than the old test. The average adjusted Kristof estimates for the new test is also .005 lower than the average adjusted Angoff-Feldt estimates for the old test.

Adjusted reliability coefficients for the new test tend to be slightly lower than adjusted reliability coefficients for the old test. The practical question to answer is "how many additional items are needed to achieve the average level of reliability that exists for the old test?" This is accomplished with a transformation of the Spearman-Brown formula (Nunnally, 1978, p. 244):

(2) $$k = r_{kk}(1-r_{11}) / r_{11}(1-r_{kk})$$

where,

k = the number the test would have to be lengthened to obtain desired reliability;

r_{kk} = desired reliability; and

r_{11} = existing reliability.

We conclude that the new test would need to be increased by a factor of 1.07, suggesting that 5 additional items would be needed in the new test to achieve a level of reliability similar to the old test.

Summary

All three methods for assessing reliability show that the new test is slightly less reliable than the old test. Internal consistency estimates show the smallest difference between the old and new test. Estimates based on parts show the largest difference.

When adjusted for a fixed standard deviation, the estimates based on covariances among parts tend to be similar to estimates based on parallel forms. Both types of estimates are lower than the internal consistency estimates.

Two factors are responsible for the decrease in reliability for the new test. First, the new test has fewer items than the old test. Second, the new test has larger item sets sharing common stimulus material than the old test, and this reduces the test's effective length. The effect of item sets is revealed by a systematic discrepancy between the internal consistency estimates, which treats items based on the same passage as independent, and the parallel-form estimates. The difference between these estimates is more pronounced for the new test than for the old test.

The new test, with longer reading passages, larger item sets, and greater emphasis on reading requires more testing time than the old test (75 minutes versus 60 minutes). The longer reading passages in the new test require more testing time and fewer questions in order to ensure test takers sufficient time to complete each test section. The trade-off between important content changes and psychometric changes needs to be taken into account when evaluating a slight decrease in test reliability for the new verbal SAT. While validity data for the new SAT are not yet available, prior research shows that the reading comprehension item has the highest validity of the verbal item types in the old test (Burton, Morgan, Lewis, & Robertson, 1989). From a validity point of view, this finding suggests that the shift toward more reading in the new test may compensate in validity gains for the slight decrease in reliability as a consequence of shortening the test's length and increasing item set size.

References

Angoff, W.H. (1953). Test reliability and effective test length. *Psychometrika, 18*, 1-14.

Burton, N.W., Morgan, R., Lewis, C., & Robertson, N.J. (1989, April). *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.

Dressel, P.L. (1940). Some remarks on the Kuder-Richardson 20 (KR-20) reliability statistic for formula scored tests. *Psychometrika, 5*, 305-310.

Feldt, L.S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education, 6*, 37-48.

Feldt, L.S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika, 40*, 557-561.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (pp. 105-146). New York: American Council on Education.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika, 39*, 491-499.

Lawrence I.M. & Schmitt, A. P. (1994). *Setting statistical specifications for the new SAT and PSAT/NMSQT* (RM-94-10). Princeton, NJ: Educational Testing Service.

Nunnally, J. (1978). *Psychometric theory*. NY: McGraw Hill.

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Wainer, H. & Thissen, D. (in press). How reliable should a test be? What is the effect of local independence on reliability? *Educational Measurement: Issues and Practice.*

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Table 1
Sample Sizes and Summary Statistics for Samples Used to Estimate Reliability from Single Test Administrations

Test	Test Form	Sample Size	Scaled Score Mean	Scaled Score SD
Old (O)	O1	2,155	405	104
	O2	3,450	401	102
	O3	2,185	403	104
	O4	1,625	433	107
	O5	2,475	422	103
	O6	3,470	426	104
	O7	3,470	425	107
New (N)	N1	3,465	439	111
	N2	3,445	441	109
	N3	3,450	439	106
	N4	3,445	442	107
	N5	3,450	427	107
	N6	3,455	425	106
	N7	3,455	428	103

Table 2
Sample Sizes and Summary Statistics for Samples Used to Estimate Parallel-Form Reliability

Test Form Pair (x,y)	Sample Size	Mean (x)	Mean (y)	SD (x)	SD(y)	Mean Diff. (y-x)	Ratio of SDs (y/x)
O1, O2	15,448	451	457	106	107	6	1.01
O4, O5	10,364	441	453	108	110	12	1.02
O1, O3	30,183	432	445	96	98	13	1.02
O4, O6	11,967	427	448	98	102	21	1.04
O2, O3	16,193	425	437	92	94	12	1.02
O4, O7	11,469	427	443	98	103	16	1.05
O5, O7	5,946	426	434	93	98	8	1.05
O5, O6	6,222	425	436	93	96	11	1.03
N1, N3	4,007	460	462	105	108	2	1.03
N4, N5	1,184	420	428	95	99	8	1.04
N1, N2	8,186	462	466	103	105	4	1.02
N1, N4	3,993	462	464	102	106	2	1.04
N3, N5	1,182	433	438	95	98	5	1.03
N2, N5	2,385	428	432	92	97	4	1.05
N2, N6	2,420	429	436	93	96	7	1.03
N2, N7	2,437	432	437	94	100	5	1.06
N3, N6	1,222	432	441	96	99	9	1.03
N3, N7	1,229	427	433	91	96	6	1.05
N1, N5	9,300	441	450	94	100	9	1.06
N4, N7	1,207	425	434	96	99	9	1.03
N1, N6	9,343	437	448	95	101	11	1.06
N1, N7	9,320	438	449	94	100	11	1.06
N4, N6	1,126	423	432	93	97	9	1.04

Note:

O = old test; N = new test

Table 3

Estimates of Reliability Based on a Single Test Administration: Seven Old Forms and Seven New Forms (National)

Test	Test Form	Var. Comp.	Var. Comp SEM	Angoff-Feldt	Kristof	Angoff-Feldt or Kristof SEM
Old (O)	O1	.919	29.59	.912	-	30.85
	O2	.922	28.51	.913	-	30.09
	O3	.923	28.86	.917	-	29.96
	O4	.925	29.32	.921	-	30.07
	O5	**.925**	**28.22**	**.923**	-	**28.58**
	O6	**.916**	**30.14**	**.904**	-	**32.22**
	O7	.922	29.87	.914	-	31.38
	Average	.922	29.22	.915	-	30.45
New (N)	N1	.925	30.41	.916	.914	32.55
	N2	**.931**	28.64	**.927**	**.923**	30.25
	N3	.930	**28.06**	.920	.921	**29.79**
	N4	.924	29.51	.913	.909	32.28
	N5	.928	28.72	.923	.919	30.45
	N6	.924	29.21	.907	.908	32.15
	N7	**.912**	**30.56**	**.898**	**.898**	**32.90**
	Average	.925	29.30	.915	.913	31.48

Notes:

1. Minimum and maximum values are shown in bold.
2. The Angoff-Feldt estimate for the new test was computed using section 1 and section 3 combined as one part, and section 2 as the other part.
3. SEMs for the old test are based on the Angoff-Feldt estimate and for the new test are based on the Kristof estimate.
4. Scaled score SEMS were obtained by multiplying the raw score SEM by the ratio of the scaled score standard deviation to the raw score standard deviation, within each sample.

Table 4

Rank Ordered Estimates of Reliability Based on Two Test Administrations

Old Test				New Test			
Test Form Pair (x,y)	Corr. (x,y)	SEM (x)	SEM (y)	Test Form Pair (x,y)	Corr. (x,y)	SEM (x)	SEM (y)
O1, O2	.918	30.35	30.64	N1, N3	.908	31.85	32.76
O4, O5	.919	30.74	31.31	N4, N5	.902	29.74	30.99
O1, O3	.895	31.11	31.76	N1, N2	.901	32.41	33.04
O4, O6	.893	32.06	33.37	N1, N4	.896	32.89	34.18
O2, O3	.892	30.23	30.89	N3, N5	.892	31.22	32.21
O4, O7	.892	32.21	33.85	N2, N5	.890	30.51	32.17
O5, O7	.887	31.26	32.94	N2, N6	.888	31.12	32.13
O5, O6	.885	31.54	32.56	N2, N7	.887	31.60	33.62
				N3, N6	.885	32.56	33.57
				N3, N7	.882	31.26	32.98
				N1, N5	.882	32.29	34.35
				N4, N7	.879	33.39	34.44
				N1, N6	.879	33.05	35.13
				N1, N7	.878	32.83	34.93
				N4, N6	.875	32.88	34.29
Average	.898	31.21	32.14		.888	31.97	33.39

Note:
Correlations are based on scaled scores for test takers who took one form of the test and took a second form of the test one or two months later.

24

Table 5

Adjusted Estimates of Reliability for a Fixed Standard Deviation (110)

Test	Test Form	Var. Comp.	Angoff-Feldt or Kristof	Average Parallel-Form
Old (O)	O1	.928	.921	.922
	O2	.933	.925	.924
	O3	.931	.926	.919
	O4	.929	.925	.917
	O5	.934	.932	.914
	O6	.925	.914	.910
	O7	.926	.919	.908
	Average	.929	.923	.917
New (N)	N1	.924	.912	.912
	N2	.932	.924	.918
	N3	.935	.927	.916
	N4	.923	.914	.912
	N5	.932	.923	.913
	N6	.929	.915	.906
	N7	.923	.911	.905
	Average	.929	.918	.912

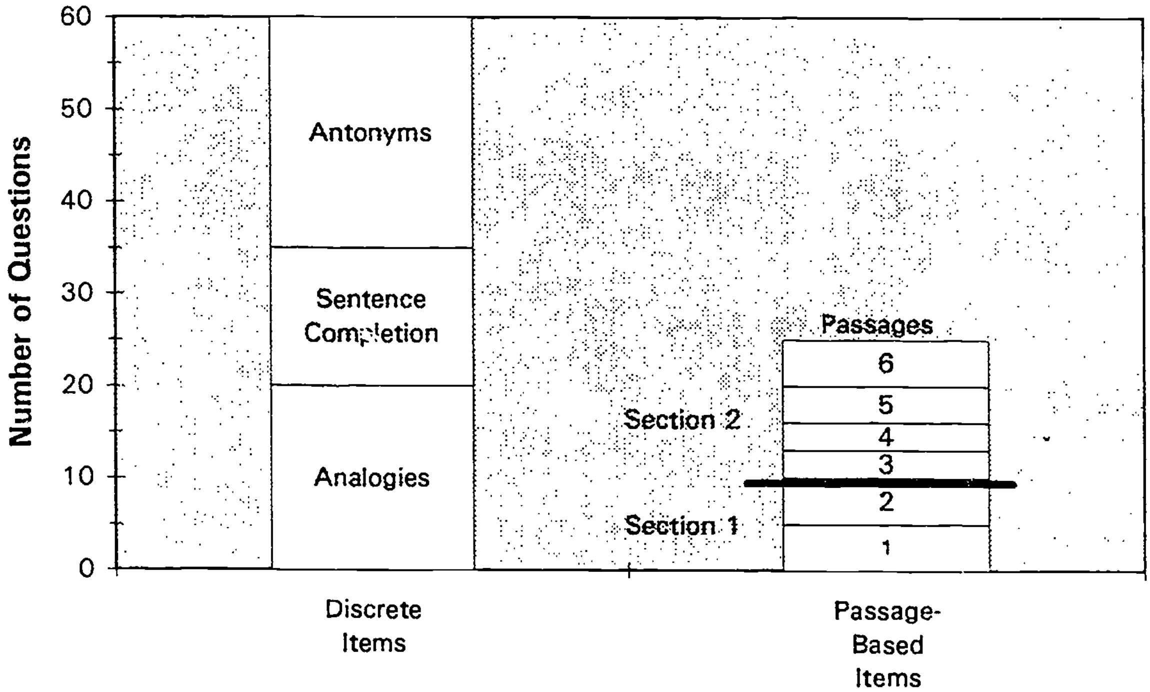
Notes:

1. Angoff-Feldt estimates were used for the old test and Kristof estimates were used for the new test.

2. Several adjusted parallel-form coefficients are available for each form (i.e., Form O1 was paired with Form O2 in one sample and with Form O3 in another sample). As a summary index for a particular form, the relevant adjusted coefficients were averaged.

Figure 1
Specified Distributions of Item Types: Old and New SAT Verbal

Old Test (85 Items)



New Test (78 Items)

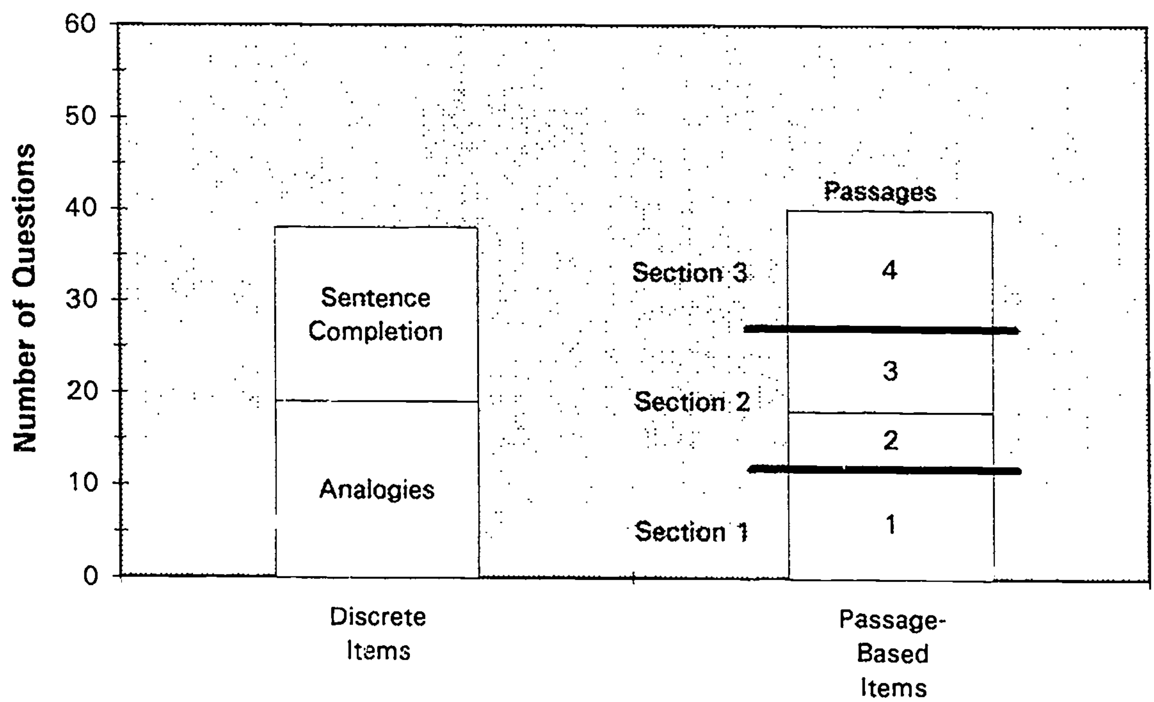
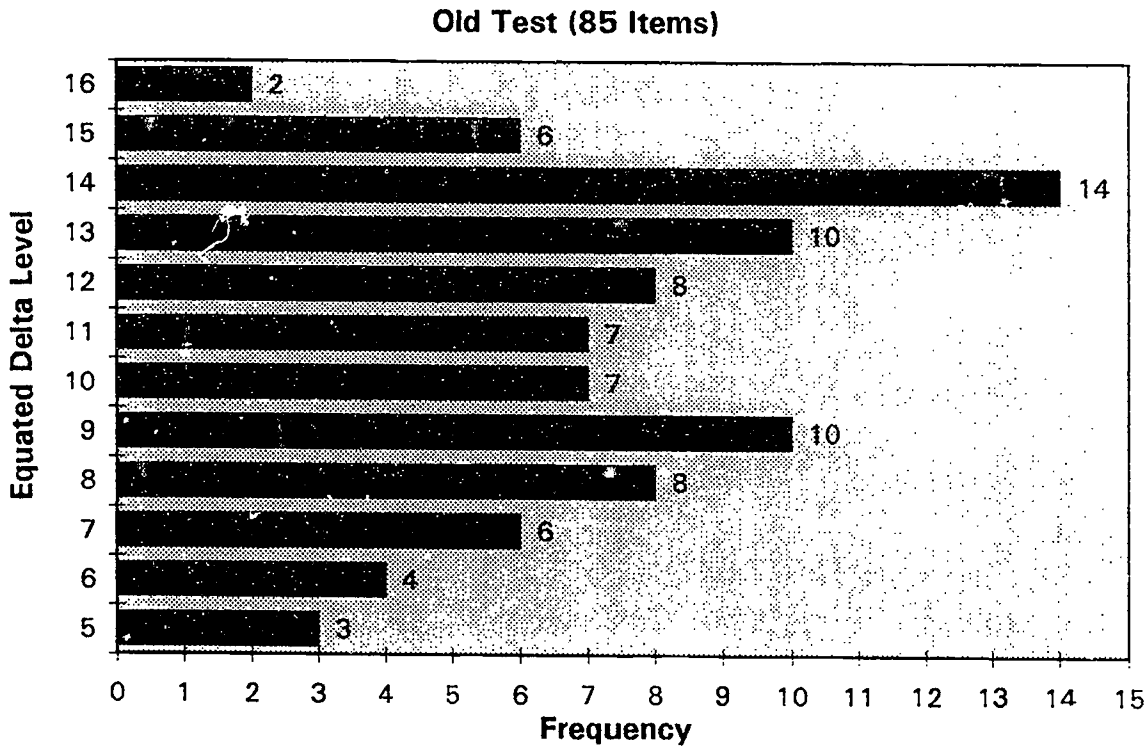


Figure 2
Specified Distributions of Item Difficulty: Old and New SAT Verbal



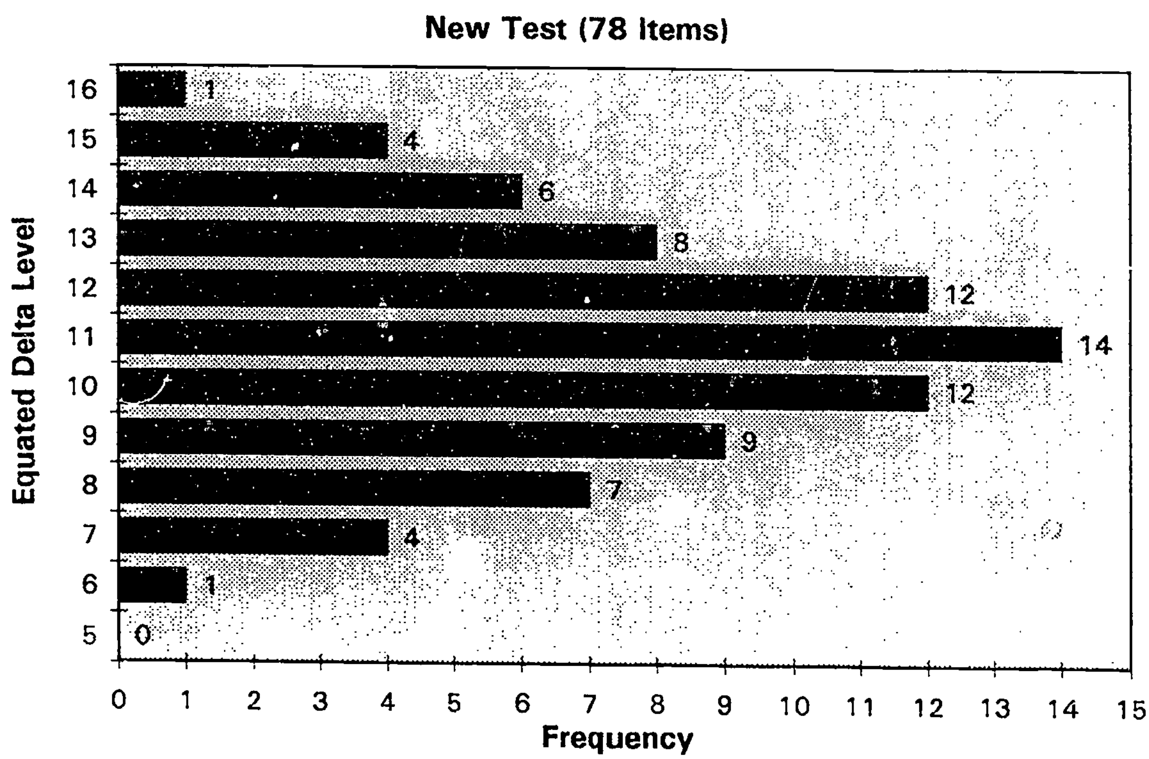


Figure 3

Formulae Used to Estimate Reliability Based on a Single Test Administration

<u>Dressel KR-20 Reliability</u>

$$reliability = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} p_i q_i' + \sum_{i=1}^{n} k^2 p_i' q_i' + 2\sum_{i=1}^{n} k p_i p_i'}{\sigma_t^2}\right],$$

where

p_i = proportion of total sample responding correctly to item i,
p_i' = proportion of total sample responding incorrectly to item i,
q_i = $1-p_i$,
q_i' = $1-q_i$,
k = correction factor for formula scoring
n = total number of items in the test section, and
σ_t^2 = variance of total formula scores for the test section.

<u>Variance-Components Reliability</u>

$$reliability = 1 - \frac{\Sigma SEM^2}{\sigma_t^2},$$

where the standard errors of measurement (SEMs) are the Dressel KR-20 SEMs for the appropriate sections and σ_t^2 is the corresponding total-score variance. SEM from Dressel KR-20 estimate =

$$SEM = \sigma_x \sqrt{1 - reliability}$$

<u>Angoff-Feldt Reliability</u>

$$reliability = \frac{cov_{12} var_T}{cov_{1T}\, cov_{2T}},$$

where

1 = 1st section,
2 = 2nd section,
var_T = $var_1 + var_2 + 2cov_{12}$,
cov_{1T} = $var_1 + cov_{12}$ and,
cov_{2T} = $var_2 + cov_{12}$

<u>Kristof Reliability</u>

$$reliability = \frac{[cov_{12}cov_{13} + cov_{12}cov_{23} + cov_{13}cov_{23}]^2}{cov_{12}cov_{13}cov_{23}var_T},$$

where

1 = 1st section,
2 = 2nd section,
3 = 3rd section and,
var_T = $var_1 + var_2 + var_3 + 2cov_{12} + 2cov_{13} + 2cov_{23}$