CollegeBoard
connect to college success™

Research Notes

Office of Research and Analysis RN-24, December 2005

Assessing the Reliability of Skills Measured by the SAT®

Maureen Ewing, Kristen Huff, Melissa Andrews, and Kinda King

Introduction

Recipients of educational score reports generally welcome the idea of receiving more descriptive feedback about examinee performance than is provided by a total score or a percentile rank indicating overall performance. This is not surprising as descriptive score reports have the potential to aid score users in the development of student-based instructional plans and/or suggest areas for classroom-based instructional intervention. In fact, under the No Child Left Behind Act of 2001 (NCLB), state testing programs are mandated by law to "produce individual student interpretive, descriptive, and diagnostic reports" [section 1111(b)(3)(c)(xii)]. Furthermore, this detailed information must be provided in such a way that the validity and reliability of the scores is maintained [section 1111(b)(3)(c)(iii)]. Due in part to this legislation, there is a general need for descriptive score reports that produce reliable scores and facilitate valid interpretations of student performance. When descriptive score reports provide reliable and valid information, they offer the possibility for improving the consequential validity of test score use and interpretation. For this reason, the College Board is committed to conducting research to identify reliable and valid ways of providing examinees with more descriptive feedback about their test performance.

In connection with the new SAT® that was introduced in March 2005, research has been under way to investigate the feasibility of providing examinees with score reports that contain feedback on skills measured by the critical reading, mathematics, and writing sections of the test. Although previous score reports for the SAT have provided cluster scores based on content specifications or item type, such scores do not typically provide great insight into whether an examinee will correctly answer a particular test item (Embretson and Gorin, 2001; Wainer, Sheehan, and Wang, 2000). This is because to meet test form assembly guidelines the content specifications for a particular domain are written to cover a range of difficulty. As a result, there are usually some easy, medium, and difficult items within each domain. An examinee of average ability would be expected to correctly answer the easy items and most of the items of medium difficulty across all content domains. In this situation, feedback based solely on content domains or item type would suggest to the student that he or she needs improvement in all areas, which is not very informative or targeted.

To generate feedback that has the potential to be more meaningful, the College Board asked subject matter experts (SMEs), including content specialists, measurement experts, and cognitive psychologists, to specify a set of skill categories hypothesized to underlie performance on each SAT test section (i.e., critical reading, mathematics, and writing). Once the models were hypothesized, items were coded to specific skill categories. Although the models that were generated allowed for the coding of items to multiple skills, the skill category that was primarily involved in solving each item was also noted. One way of providing feedback to examinees is to report skill scores based on the primary skill codes generated by the SMEs, which are described in Table 1. Such skill scores have the potential to be more informative than conventional cluster scores based on content or item type because the skills reflect an underlying model of student performance in the domain.[1]

[1] For a detailed description of the theoretical framework used by the SMEs to specify the skill categories, as well as the processes used to code items to those categories, see Huff (2004), O'Callaghan, Morley, and Schwartz (2004), and VanderVeen (2004).

Table 1

Skills Measured by the SAT Critical Reading, Mathematics, and Writing Tests

SAT Test Section	*Skill Number*	*Skill Description*
Critical Reading	Reading_Sk1	Determining word meaning
	Reading_Sk2	Understanding sentences
	Reading_Sk3	Understanding larger sections of text
	Reading_Sk4	Analyzing purpose, audience, and strategies
Mathematics	Math_Sk1	Applying basic mathematics knowledge
	Math_Sk2	Applying more advanced mathematics knowledge
	Math_Sk3	Managing complexity
	Math_Sk4	Modeling and insight
Writing	Writing_Sk1	Managing word choice and grammatical relationships between words
	Writing_Sk2	Managing grammatical structures used to modify or compare
	Writing_Sk3	Managing phrases and clauses in a sentence
	Writing_Sk4	Managing order and relationships of sentences and paragraphs

Note: These skills remain under development and should not be considered final.

There are psychometric challenges, however, to reporting skill scores in this manner. One challenge concerns the extent to which the skill scores are reliable across alternate test forms. This is a challenge because it cannot be assumed that the test specifications for the SAT would incorporate skill-level requirements. As a result, it is necessary to evaluate the extent to which skill scores are influenced by form-to-form differences in skill coverage. This research sought to address these issues by evaluating the reliability of skill scores in the situation where test specifications were not changed to directly incorporate the skills. Both the alternate-form reliability and internal consistency of the skill scores were estimated. Rater reliability (i.e., how consistently SMEs code items to skills) was not addressed in this study; however, rater reliability was investigated by Gierl, Leighton, Wang, and Tan (2005) for one of the forms used in this study. According to their results, the generalizability coefficient for critical reading was .91 and the generalizability coefficient for mathematics was .98, both of which are indicative of high-rater agreement. In the case of critical reading, the variance component for raters accounted for only .16 percent of the total variance and, in the case of mathematics, it accounted for only .15 percent of the total variance, which is further indicative of high-rater agreement. Writing was not investigated.

Method

Design

Two forms of the SAT critical reading, mathematics, and writing tests were used for this study. Form 1 was developed for the spring 2003 new SAT field trial and Form 2 was developed as a practice test for the new SAT study guide. For writing, current test specifications were not finalized in time for this study. As a result, the writing tests used in this study were each composed of one section of 37 multiple-choice items. The essay section was not included because a detailed scoring rubric was developed to explain the meaning of essay scores.[2] To minimize testing time, schools were asked to administer two forms of the critical reading, mathematics, or writing tests, as opposed to being asked to administer two forms of an entire SAT. Separate test booklets were developed for each test, and the order of the test form was counterbalanced within each booklet so factors such as fatigue, motivation, and practice effects could be controlled.

Recruitment

The target sample was composed of high school juniors not requiring special testing accommodations. High school seniors were not targeted because data collection occurred in February and March and, as a result, there was concern about their motivation to perform well. In addition, the target sample was expected to be reasonably representative of key demographic characteristics of the 2003 College-Bound Seniors (CBsrs) cohort, including student ethnicity, school type, and school location. The target sample size for each of the three SAT tests was 500 students, which was chosen because it exceeded our minimum sample size requirements for correlation analyses (i.e., $n = 300$) evaluated by using Fisher's z formula to obtain 90 percent and 95 percent confidence intervals for population correlations that we expected to range from .60 to .80. To ensure adequate sample sizes, students were over-recruited at a rate of approximately 40 percent; however, final sample sizes fell somewhat short of the target, as will be seen following.

Recruitment began in the fall of 2003 when approximately 1,000 letters were sent to high schools around the country. The list of potential schools from which to recruit was obtained from Educational Testing Service and included all schools that were originally invited to submit an application to participate in the spring 2003 field trial of the new SAT. Our recruitment strategy was to focus on those schools

[2] The current writing test is composed of two multiple-choice sections (35 minutes) and one essay section (25 minutes).

that expressed interest in the original field trial. As a whole, these schools were not necessarily representative of all SAT schools. However, because our target sample size was small, we expected this approach to yield a sufficient number of schools from which to select those with the most desirable demographic characteristics.

The recruitment letter described the purpose of the study and the requirements for participation. In addition, it informed schools that only high school juniors not requiring special testing accommodations were eligible to participate. It also emphasized that because the study was seeking a cross section of the student population, the sample of test-takers should be representative of a school's entire junior class and not solely composed of, for example, their Advanced Placement Program® students or their best English and math students.

A total of 187 schools expressed interest in participating. Schools identified the total number of students they anticipated testing as well as the ethnic distribution of those potential participants. In addition, schools indicated which of the three subject areas they were willing to administer. They were able to mark one, two, or all three. To guide the selection of schools, the demographic characteristics of the 2003 CBsrs, namely, student ethnicity, school type (i.e., public versus private), and region of the country. The regions are defined by the College Board and, thus, may not directly correspond to regions defined geographically. See the Appendix for a list of states by College Board regions. The combination of schools for each subject area that came closest to matching the demographic characteristics of the 2003 CBsrs cohort were invited to participate. When selecting the final set of schools, it became necessary to give more emphasis to student ethnicity and school type than to school location due to small target sample sizes.

Sample

In total, 16 schools participated. All schools administered the critical reading, mathematics, or writing tests, except for one school that administered both the writing and critical reading tests to separate students. Across the participating schools, seven administered critical reading, five administered mathematics, and five administered writing. A total of 1,696 students from these schools returned answer sheets; however, not all students returned usable data. The steps that were taken to clean the data are described next.

Students were removed from the analyses for one of three reasons: (1) failure to take both forms of the critical reading, mathematics, or writing tests; (2) failure to provide a booklet code; or (3) not appearing motivated to perform well. A total of 451 students were removed based on these criteria. The most frequent reason for removal was missing booklet codes. For example, at one high school alone, 300 students returned answer sheets for the writing portion of the study without booklet codes. Missing booklet codes made it impossible to score the responses. Unmotivated examinees were operationalized as those with a large number of missing responses (e.g., an entire test section) and/or those with obvious pattern markings. Examples of obvious pattern markings included a string of all A's, or B's, or a continuous pattern such as ABCDE. To identify unmotivated examinees, the raw response patterns of low-ability examinees were reviewed. Low-ability examinees were defined as those who answered only 20 percent or fewer of the items correctly on either form.

Despite the request that only eleventh-graders participate, analysis of self-reported grade-level data indicated that 153 ninth-, tenth-, and twelfth-graders participated. In fact, one writing school tested only ninth- and tenth-graders. In addition, six students did not report a grade level. Rather than removing these students, several analyses were conducted to compare performance of the total group to juniors only. The results of these analyses demonstrated that the groups were reasonably similar and, therefore, this report focuses on the results obtained for all students. The final sample sizes for critical reading, mathematics, and writing were 494, 485, and 260 students, respectively.

Table 2 displays the percentage of examinees by gender, ethnicity, school type, and school location for critical reading, mathematics, and writing separately. The final column shows the same information for the 2003 CBsrs cohort. The table shows that the gender and ethnicity of the sample match the target population (i.e., CBsrs 2003) reasonably well, although less well for writing. In terms of school type, a majority of students were from public schools for the mathematics (67 percent) and writing (73 percent) portions of the study, as desired. For critical reading, approximately half of the participants were from public high schools (51 percent) and half were from nonpublic high schools (49 percent). In terms of school location, no schools participated from the Midwest; however, all other regions participated in at least one portion of the study. As can be seen, a majority of the

Table 2

Sample Characteristics

	Critical Reading	*Mathematics*	*Writing*	*CBsrs 2003*
Gender				
Female	53%	48%	60%	54%
Male	47%	52%	40%	46%
Ethnicity				
African American	11%	11%	6%	12%
American Indian	1%	.2%	.4%	1%
Asian American	11%	15%	17%	10%
Hispanic	6%	10%	6%	10%
White	67%	60%	63%	64%
Other	5%	5%	6%	4%
School Type				
Public	51%	67%	73%	83%
Nonpublic	49%	33%	27%	17%
CB Region				
New England	14%	—	4%	9%
Middle States	45%	29%	94%	28%
Midwest	—	—	—	10%
South	17%	39%	—	22%
Southwest	15%	12%	3%	10%
West	9%	20%	—	21%

Note: Due to rounding error and missing data, percentages may not always sum to 100 percent.

students who participated in the writing portion of the study were from the Middle States region. This is due to the fact that one school in the Midwest that was set to administer the writing portion withdrew shortly before testing was scheduled to begin (originally 17 schools agreed to participate). The school that submitted 300 answer sheets for the writing test without booklet codes was from the Southwest.

Procedure

Participation in this study required schools to administer two forms of the SAT critical reading, mathematics, and writing tests between the period of February and mid-March 2004. To encourage schools to participate, two possible administration plans were permitted in order to accommodate a school's schedule as best as possible. The two options were same-day testing or multiday testing. If schools selected the multiday testing option, they were required to administer an entire test from beginning to end during one sitting; however, requirements as to the time between administrations of each form were not stipulated as long as testing was completed by mid-March. Ten schools followed the same-day administration plan and the remainder followed the multiday administration plan. Schools were asked to provide exact dates of testing so that the duration of time between testing could be estimated; however, very few schools did so. As a result, it was not possible to determine the exact amount of time that elapsed between testing for those schools that followed a multiday administration plan.

The total time allotted for the critical reading and mathematics portions of the study was three hours. This included 140 minutes for actual testing (70 minutes per form), 20–25 minutes for administration of materials and completion of the answer sheet, and 10 minutes for a break between administrations of each form (assuming a one-day administration schedule). The total time allotted for the writing portion of the study was 90 minutes. This included 60 minutes for actual testing (30 minutes per form), as well as time for the administration of test materials, completion of the answer sheet, and a 10-minute break. The section time for the writing test in the new SAT field trial was 25 minutes. For this study, we extended time by five minutes, as field-trial analyses showed that the writing test was speeded (meaning that the time limit was not long enough to allow most examinees to finish the test).

Test materials (e.g., test booklets, answer sheets, preaddressed mailing labels, and test administration instructions) were sent to schools approximately one week prior to the date selected for testing. As previously discussed, the order of the test form (i.e., Form 1 first or Form 2 first) was counterbalanced so factors such as fatigue, motivation, and practice effects could be controlled. For example, if a student received math booklet A, the student took math Form 1 first and math Form 2 second, whereas a student who received math booklet B took the forms in the reverse order. Prior to distribution, test booklets were spiraled so that each school received about an equal number of A and B booklets. The test administration instructions provided detailed information about: (1) what to tell students prior to testing; (2) the length of time needed for testing; (3) how to start and stop testing; (4) how to distribute test booklets; and (5) what to do at the end of testing. Upon receipt of test materials, schools were asked to carefully review the test administration instructions and check to make sure they had received all required materials.

Analyses

Prior to conducting the substantive analyses for this study, statistical analyses were conducted to ensure that the coun-

terbalancing controlled for unwanted motivation, fatigue, and practice effects. To conduct these analyses, the critical reading, mathematics, and writing number correct scores were summed across both forms. For example, a student's number correct score on mathematics Form 1 was added to his or her number correct score on mathematics Form 2. This process generated a combined total score for each critical reading, mathematics, and writing for every student. The expectation was that within critical reading, mathematics, or writing the combined total scores should be about equal for students who were administered booklet A (i.e., Form 1 followed by Form 2) compared to those who were administered booklet B (i.e., Form 2 followed by Form 1).

To evaluate the reliability of the skill scores, two estimates were computed: (1) internal consistency and (2) alternate-form reliability. Internal consistency was estimated separately for each skill by form using the formula for Cronbach's alpha, as shown below:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right) \qquad (1)$$

where k is the number of items comprising the skill, σ_i^2 is the variance of item *i*, and σ_x^2 is the total skill variance. Coefficient alpha ranges from zero to one, with values closer to one indicating that the items are more highly intercorrelated with one another. In other words, higher values indicate that the items are measuring the same underlying construct.

To estimate the alternate-form reliability of each skill, Pearson-product moment correlations were estimated within skill across form; that is, for each skill, the raw number correct score on Form 1 was correlated with the corresponding raw number correct score on Form 2 using the formula presented below:

$$r_{xy} = \frac{\sum (x-\mu_x)(y-\mu_y)}{N\sigma_x\sigma_y} \qquad (2)$$

Where μ_x and μ_y are the respective average scores for a particular skill on Form 1 and Form 2, *N* is the number of students, and σ_x and σ_y are the respective standard deviations for the skill on Form 1 and Form 2. Alternate-form reliability estimates typically range from zero to one with higher values indicating that students perform about the same on both forms.

Results

Counterbalancing Check

To evaluate the assumption that the two groups were randomly equivalent, independent sample *t*-tests were conducted where the independent variable was order of administration and the dependent variable was the combined total score for critical reading, mathematics, or writing. For this study, the maximum combined total score that is possible for critical reading, mathematics, and writing is 134, 108, and 50, respectively. Table 3 shows the mean scores and standard deviations (SD) for each SAT section by order of administration. The results for critical reading, t (492) = -.563, p = .574, mathematics, t (483) = 1.240, p = .216, and writing, t (264) = .469, p = .639, were not statistically significant.

Cohen's effect size, *d*, was calculated to determine the magnitude of the mean score difference between students who were administered booklet A versus students who were administered booklet B. Cohen's effect size, *d*, describes score differences in terms of standard deviation units. Conventionally, a value of at least .20 is considered to be a small effect size, a value of at least .50 is considered a medium effect size, and a value of at least .80 is considered a large effect size (Cohen, 1988). To compute Cohen's *d*, the following formula was used where M_A refers to the mean combined total score for students who took booklet A, M_B refers to the mean combined total score for students who took booklet B, and σ equals the pooled standard deviation for both groups.

$$d = M_A - M_B \,/\, \sigma \qquad (3)$$

Our analyses found an effect size of .05 for critical reading, .11 for mathematics, and .06 for writing, all of which fall well below Cohen's criteria for even a small effect size. In summary, the results from the *t*-test and the effect-size analyses suggest that the counterbalancing controlled for unwanted effects, and indicate that the two groups of students are randomly equivalent.

Table 3

Means and Standard Deviations for Scores Combined Across Form 1 and Form 2

SAT Section	*Booklet A Mean Score (SD)*	*Booklet B Mean Score (SD)*
Critical Reading	64.78 (26.32)	63.45 (25.83)
Mathematics	58.00 (20.17)	55.71 (20.50)
Writing	36.46 (12.35)	37.18 (12.56)

Descriptive Results

Table 4 reports the mean, median, standard deviation (SD), standard error of measurement (SEM), and the total number of items that were coded to each skill category for critical reading, mathematics, and writing, respectively. Notice that for writing, the number of items per skill on Form 1 and Form 2 are not well distributed. Writing_Sk4 on Form 1 is especially problematic with only three items coded to it, and will not be examined further in this report.

Reliability

Before discussing the internal consistency estimates at the skill level, it is worth noting that the internal consistency estimates at the total test level varied by SAT section (i.e., critical reading, mathematics, and writing), but *not* by form (i.e. Form 1 versus Form 2). For both Form 1 and Form 2, the internal consistency estimates at the total test level were .93 for critical reading, .92 for mathematics, and .83 for writing.

Table 5 displays the internal consistency estimates at the skill level for both forms. Internal consistency estimates range from .69 to .84 for critical reading, .68 to .81 for mathematics, and .40 to .67 for writing. The internal consistency estimates are lowest for Writing_Sk2, which was measured by the fewest items on both forms. While most of the internal consistency estimates are reasonably good, it is important to emphasize that the interpretation of reliability estimates can be subjective. At the total test level, acceptable internal consistency estimates usually fall above .85. For skill-level feedback, however, no clear-cut guidelines exist for judging acceptable levels.

With respect to alternate-form reliability, estimates at the total test level are .88 for critical reading, .91 for math-

Table 5

Internal Consistency and Alternate-Form Reliability Estimates by Skill

Skill	*Internal Consistency Form 1*	*Internal Consistency Form 2*	*Alternate-Form Reliability*
Reading_Sk1	.69	.60	.65
Reading_Sk2	.84	.82	.77
Reading_Sk3	.82	.87	.79
Reading_Sk4	.69	.68	.60
Math_Sk1	.80	.82	.78
Math_Sk2	.77	.79	.78
Math_Sk3	.68	.80	.71
Math_Sk4	.81	.64	.72
Writing_Sk1	.64	.68	.68
Writing_Sk2	.40	.45	.44
Writing_Sk3	.67	.56	.56

Table 4

Descriptive Information by Skill Across Forms

	Reading_Sk1		*Reading_Sk2*		*Reading_Sk3*		*Reading_Sk4*	
	Form 1	*Form 2*	*Form 1*	*Form 2*	*Form 1*	*Form 2*	*Form 1*	*Form 2*
No. Items	12.00	7.00	24.00	18.00	21.00	31.00	10.00	11.00
Mean	6.05	3.17	12.13	9.53	10.50	14.66	3.77	4.30
Median	6.00	3.00	12.00	10.00	10.00	14.00	3.00	4.00
SD	2.64	1.73	5.13	4.05	4.68	6.67	2.43	2.60
SEM	1.47	1.09	2.05	1.72	1.98	2.40	1.36	1.47
	Math_Sk1		*Math_Sk2*		*Math_Sk3*		*Math_Sk4*	
	Form 1	*Form 2*	*Form 1*	*Form 2*	*Form 1*	*Form 2*	*Form 1*	*Form 2*
No. Items	17.00	15.00	11.00	11.00	8.00	14.00	18.00	14.00
Mean	13.31	11.33	6.93	7.15	3.83	4.51	5.62	4.18
Median	14.00	12.00	7.00	8.00	4.00	4.00	5.00	4.00
SD	3.26	3.31	2.86	2.86	2.16	3.26	3.77	2.54
SEM	1.46	1.40	1.37	1.31	1.22	1.46	1.64	1.52
	Writing_Sk1		*Writing_Sk2*		*Writng_Sk3*		*Writing_Sk4*	
	Form 1	*Form 2*	*Form 1*	*Form 2*	*Form 1*	*Form 2*	*Form 1*	*Form 2*
No. Items	13.00	15.00	7.00	5.00	13.00	10.00	3.00	6.00
Mean	6.36	7.67	3.09	2.70	6.84	4.99	1.05	3.29
Median	6.50	8.00	3.00	3.00	7.00	5.00	1.00	4.00
SD	2.62	3.03	1.54	1.36	2.77	2.11	0.96	1.67
SEM	1.57	1.72	1.19	1.01	1.59	1.40	0.73	1.07

ematics, and .80 for writing. At the skill level, the alternate-form reliability estimates range from .60 to .79 for critical reading, .71 to .78 for mathematics, and .44 to .68 for writing (see Table 5). As was the case with estimates of internal consistency, the alternate-form reliability estimates for writing are notably lower than those for critical reading or mathematics.

Discussion

The main purpose of this study was to estimate the reliability of the SAT skill scores generated by SMEs. Findings showed that with the exception of the writing skills, most skills exhibited acceptable internal consistency and alternate-form reliability estimates. As mentioned, interpretation of reliability estimates, especially at the skill level, may be subjective as no clear-cut guidelines exist for judging acceptable levels. One way to place these findings into context is to compare them to the subscore reliabilities reported for other large-scale tests. For example, ACT reports seven subscores in the areas of English, Mathematics, and Reading. The internal scale score reliability estimates for these subscores, averaged across five administrations, was reported to range from .71 to .85 (ACT, 1997, p.32). With the exception of writing, the estimates reported by ACT are generally similar to the internal consistency and alternate-form reliability estimates obtained for the skills investigated in this study.

The limitations of this study are important to mention. The first concerns the writing portion of the study. It had the smallest sample size and was composed of approximately 40 percent of students who were not invited to participate (i.e., freshmen, sophomores, seniors) or who chose not to report a grade level. In addition, the writing sections that were administered corresponded to test specifications that have since been modified. Future work should therefore reexamine the reliability of the writing skills. A second shortcoming of this study relates to the data collection process. Despite seemingly clear test administration instructions, schools did not always follow guidelines. Some schools allowed freshmen, sophomores, and/or seniors to participate. Other schools did not make sure that students documented their booklet code on the answer sheet. Because our target sample size was relatively small, these problems had a large impact. If we were to conduct future studies of this kind, one recommendation would be to ask proctors to sign a good-faith "contract" that explicitly outlines roles and expectations. Finally, the target sample was not well met in terms of the distribution of participants by school type and school location (defined by College Board regions). As a consequence, the results may not be entirely representative of the original target sample. In addition, these results may not fully generalize to an operational setting. Because some students may have been less motivated to perform well than other students, the alternate-form reliability estimates may be somewhat inflated due to a greater performance distinction between motivated and nonmotivated examinees.

While this study investigated the reliability of individual skills, future research might investigate the similarity of the entire skill profile. Cronbach and Gleser (1953) were the first to discuss the notion of assessing the similarity of profiles and, recently, Brennan (2005) expanded on the test theory to support such analyses. Future research may also explore alternative methods for providing detailed, skill-based feedback to examinees. Promising areas include recent modifications to cognitive diagnosis methodologies, such as the attribute hierarchy method (Leighton, Gierl, and Hunka, 2004) or feedback based on distractor analyses (Luecht, 2005).

Although this study showed that most skills exhibited acceptable internal consistency and alternate-form reliability estimates, it is important to mention that this study did not address the validity of the skill scores. Additional research is needed to show that the skills actually measure what they purport to measure, and that any score reports based on the skill scores support valid inferences about examinee performance.

Maureen Ewing is an assistant research scientist at the College Board; Kristen Huff is director of K–12 research at the College Board; and Melissa Andrews and Kinda King are policy analysts at the College Board.

References

ACT (1997). *ACT assessment: Technical manual.* Iowa City, Iowa.

Brennan, R. L. (2005). *Some test theory for the reliability of individual profiles*. (CASMA Research Report Number 12). Iowa City, Iowa: University of Iowa, Center for Advanced Studies in Measurement and Assessment.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing the similarity between profiles. *The Psychological Bulletin*, *50*, 456–73.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343–68.

Gierl, M. J., Leighton, J. P., Wang, C., & Tan, X. (2005). *Technical Report #2: Evaluating primary skill categories*. Unpublished technical report. New York: The College Board.

Huff, K. (2004). *A practical application of evidence centered design principles: Coding items for skills*. In K. Huff (Organizer), Connecting curriculum and assessment through meaningful score reports. Symposium conducted at the meeting of the National Council on Measurement in Education, San Diego.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement*, *41*, 205–37.

Luecht, R. M. (2005, April). *Extracting multidimensional information from multiple-choice question distractors for diagnostic scoring*. In Mark Gierl and K. Huff (Organizers), Enhancing the diagnostic value of large-scale achievement tests: Technical developments and applications. Symposium conducted at the meeting of the National Council on Measurement in Education, Montreal.

No Child Left Behind Act of 2002. Retrieved November 21, 2005, from http://www.ed.gov/policy/elsec/leg/esea02/pg2.html#sec1111.

O'Callaghan, R., Morley, M., & Schwartz, A. (2004). *Developing skill categories for the SAT math section*. In K. Huff (Organizer), Connecting curriculum and assessment through meaningful score reports. Symposium conducted at the meeting of the National Council on Measurement in Education, San Diego.

VanderVeen, A. (2004). *Toward a construct of critical reading for the new SAT*. In K. Huff (Organizer), Connecting curriculum and assessment through meaningful score reports. Symposium conducted at the meeting of the National Council on Measurement in Education, San Diego.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–40.

Appendix: States by College Board Region

States by College Board Region

West	*Southwest*	*Midwest*	*South*	*Middle States*	*New England*
Alaska	Arkansas	Illinois	Alabama	Delaware	Connecticut
Arizona	New Mexico	Indiana	Florida	Dist. of Columbia	Maine
California	Oklahoma	Iowa	Georgia	Maryland	Massachusetts
Colorado	Texas	Kansas	Kentucky	New Jersey	New Hampshire
Hawaii		Michigan	Louisiana	New York	Rhode Island
Idaho		Minnesota	Mississippi	Pennsylvania	Vermont
Montana		Missouri	North Carolina		
Nevada		Nebraska	South Carolina		
Oregon		North Dakota	Tennessee		
Utah		Ohio	Virginia		
Washington		South Dakota			
Wyoming		West Virginia			
		Wisconsin			

Office of Research and Analysis
The College Board
45 Columbus Avenue
New York, NY 10023-6992
212 713-8000