Illinois Standards Achievement Test: Writing

2008 Technical Manual

Illinois State Board of Education
Assessment Division

Table of Contents

1. PURPOSE AND DESIGN OF THE ISAT TESTING PROGRAM

In Spring 2008, students in grades 5, 6, and 8 took Illinois Standards Achievement Tests (ISAT) in writing. Approximately 450,000 students enrolled in schools across the state participated in the testing program. ISAT measures the extent to which students are meeting the Illinois Learning Standards. Illinois teachers and curriculum experts developed the ISAT tests in cooperation with the Illinois State Board of Education (ISBE).

This manual provides technical information about the 2008 tests and test administration. It describes the tests and assessment approaches and addresses technical concerns.

Test Development

Each ISAT test is designed to ensure that its results validly and fairly assess the Illinois Learning Standards. The selection of items and assembly of each test is guided by a set of specifications: The Illinois Assessment Frameworks[1]. These specifications were developed by Illinois educators to help ensure that test content corresponds to the purposes, objectives, and skills framed by the learning standards and to define those elements of the standards that are suitable for state testing.

Illinois teachers and administrators participate in all phases of the test development process: item writing, item selection, bias review, and test assembly. The State Board of Education convenes a series of advisory committees to ensure that test development is continually informed and guided by the recommendations of content authorities, measurement specialists, and practitioners. The following evaluation criteria are applied to all assessment material used in the Illinois program:

> *Content.* Every item is screened for alignment with the Assessment Frameworks, grade-level appropriateness, importance, and clarity. Incorrect choices (for multiple-choice items) are reviewed for plausibility. In tests other than reading, the complexity of the text of the questions is kept to the minimum necessary to state the problem.
>
> *Difficulty.* Items are field tested on large samples of students prior to their inclusion in tests to develop a statistical profile for each item. Items that are too easy or too difficult and, therefore, provide little or no information are omitted.

[1] http://www.isbe.net/assessment/pdfs/IAFWriting.rtf

Precision. Point-biserial (i.e., item-test) correlations evaluate the extent to which an item distinguishes between less proficient and more proficient students. Reviewers usually omit items with a point-biserial of less than .30 and select items with the highest point-biserial.

Fairness. Test items and forms undergo regular sensitivity reviews and statistical analyses to ensure that all materials meet fairness criteria with respect to the cultural and ethnic diversity of Illinois public schools.

ISBE takes several precautions to help ensure test security. Test materials shipped to schools are packaged and sealed. The administration of tests is standardized. A series of manuals provides guidance on security and other issues to the district testing coordinator, school testing coordinator, and classroom test administrator. After administration, all materials are removed from schools and returned to a central facility for processing and secure destruction of unneeded materials.

The state goal for writing states that the student will be able to write standard English in a grammatical, well-organized, and coherent manner for a variety of purposes. The ability to write clearly is essential to any person's effective communication. Students with high-level writing skills can produce documents that show planning and organization and effectively convey the intended message and meaning.

The learning standards associated with the goal are as follows:

3A. Use correct grammar, spelling, and punctuation.

3B. Compose well-organized and coherent writing.

3C. Communicate ideas in writing to accomplish a variety of purposes.

The writing assessment uses three types of prompts, which represent persuasive, expository, and narrative discourse modes. Persuasive topics require students to take a position on an issue or to state a problem and solution. Expository topics require students to explain, interpret, or describe something objectively and clearly. Narrative topics require students to reflect upon and describe an experience or event from personal knowledge. Readers evaluate each paper with respect to its focus, support/elaboration, organization, and conventions. They also evaluate how effectively the paper integrates these features.

Students in grade 5 wrote one assigned essay on an expository topic. Students in grades 6 and 8 wrote one persuasive and one narrative essay. All students within a grade received the same assignment.

Readers score all papers with respect to four specific features (focus, support/elaboration, organization, and conventions) and a holistic feature (integration). Descriptions of these features follow:

- *Focus:* the degree to which the subject, issue, theme, or unifying event of the composition is clear and maintained.

- *Support/Elaboration:* the quality of the detail or support through reasons and explanations.

- *Organization:* the extent to which a clear structure or plan of development is maintained and the points logically related to each other and the text structure.

- *Conventions:* the extent to which the writer demonstrates adequate knowledge of standard English.

- *Integration:* the extent to which the paper as a whole uses the four features (focus, support/elaboration, organization, and conventions) to address the assignment.

Readers rate a paper's first three features and its overall integration on a scale from 1 (absent) to 6 (well developed). The conventions feature is evaluated on a scale from 1 (little or no discernable knowledge of conventions) to 3 (strong knowledge of conventions demonstrated). A composite writing score is derived from the raw feature scores according to the following formula:

Focus + Support/Elaboration + Organization + Conventions + (2 x Integration)

The overall writing score ranges from 6 to 33. For students who wrote more than one essay (grades 6, 8), writing scores for each essay were averaged and then rounded up. Thus, individual student scores at all grades are reported as whole numbers. Scores for schools, districts, and the state are reported to one decimal place.

Item Bias Review and DIF Analysis

All ISAT items are screened for potential bias by teacher panels, administrators, and vendor content experts. They are checked during three stages: item writing, item review, and data review. First, all of the teachers who are involved in item writing are trained and instructed to balance ethnic and gender references and to avoid gender and ethnic stereotypes. Then, another group of teachers is invited to the item review meetings to screen for potential language and content bias. Items approved by the item review committee are field tested and analyzed for differential item functioning. Last, Illinois administrators, vendor content experts, and a group of teachers review each item based on statistical inputs in data review meetings.

Differential item functioning (DIF) refers to the different statistical properties of an item between groups. ISAT DIF analyses are done in three ways: males versus females, White versus Black, and White versus Hispanic. The statistical method used for writing is the polytomous extension of the Mantel-Haenszel procedure.

Its expression is

$$MH - \chi^2 = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - .5\right]^2}{\sum_m Var(R_{rm})},$$

where R_{rm} is the actual number of reference-group scores at score point *m*, $E(R_{rm})$ is the expected number of reference-group score at score point *m*, and $Var(R_{rm})$ is the variance of R_{rm}.

Evaluation of DIF severity follows the ETS DIF categories, A, B, and C, where A represents a negligible DIF, B represents a moderate DIF, and C represents a large DIF.

Table 1.1 presents results of the DIF tests for the 2008 tests. The χ^2 value tests the null hypothesis that there is no DIF. The value of SMD represents the proportion of a score point by which the focal and reference groups differ, after adjusting for group differences in the distribution of the matching variable, and z_{SMD} is equal to SMD divided by its standard error. As Table 1.1 shows DIF is negligible for the writing scales.

Table 1.1
DIF Analysis Results

Writing Element	Mantel-Haenszel Chi-Square	df	p	Reference Group N	Focus Group N	Standardized Mean Difference (SMD)	Standard Error of SMD	Z(SMD)	p(Z)	Flag
Grade 5: Male/Female Comparisons										
Focus	8.40	1	0.00	3518	3450	0.03	0.01	3.06	0.00	A
Support	0.91	1	0.34	3518	3450	-0.01	0.00	-1.44	0.15	A
Organization	7.01	1	0.01	3518	3450	-0.01	0.00	-2.69	0.01	A
Conventions	0.89	1	0.35	3518	3450	-0.01	0.01	-0.84	0.40	A
Integration	1.83	1	0.18	3518	3450	-0.01	0.01	-1.45	0.15	A
Grade 5: White/Black Comparisons										
Focus	0.00	1	0.95	3970	1375	0.00	0.01	0.22	0.83	A
Support	5.97	1	0.01	3970	1375	0.01	0.01	1.87	0.06	A
Organization	1.65	1	0.20	3970	1375	0.01	0.01	1.73	0.08	A
Conventions	5.03	1	0.02	3970	1375	-0.02	0.01	-1.98	0.05	A
Integration	0.42	1	0.52	3970	1375	-0.02	0.01	-1.67	0.10	A
Grade 5: White/Hispanic Comparisons										
Focus	5.16	1	0.02	3970	1136	-0.02	0.01	-2.25	0.02	A
Support	7.13	1	0.01	3970	1136	0.02	0.01	2.92	0.00	A
Organization	0.83	1	0.36	3970	1136	0.00	0.01	-0.25	0.80	A
Conventions	0.81	1	0.37	3970	1136	0.01	0.01	1.09	0.28	A
Integration	0.51	1	0.48	3970	1136	0.01	0.01	1.07	0.28	A
Grade 6: Male/Female Comparisons										

Writing Element	Mantel-Haenszel Chi-Square	df	p	Reference Group N	Focus Group N	Standardized Mean Difference (SMD)	Standard Error of SMD	Z(SMD)	p(Z)	Flag
Focus	13.69	1	0.00	3701	3458	0.03	0.01	3.60	0.00	A
Support	0.00	1	0.99	3701	3458	0.00	0.01	-0.19	0.85	A
Organization	9.77	1	0.00	3701	3458	-0.02	0.01	-2.94	0.00	A
Conventions	3.18	1	0.07	3701	3458	0.02	0.01	2.00	0.05	A
Integration	2.40	1	0.12	3701	3458	0.00	0.01	-0.08	0.93	A
Grade 6: White/Black Comparisons										
Focus	0.02	1	0.88	3897	1502	0.01	0.01	0.51	0.61	A
Support	8.45	1	0.00	3897	1502	0.03	0.01	3.17	0.00	A
Organization	6.54	1	0.01	3897	1502	0.02	0.01	2.87	0.00	A
Conventions	40.51	1	0.00	3897	1502	-0.09	0.01	-6.20	0.00	A
Integration	9.73	1	0.00	3897	1502	-0.03	0.01	-2.81	0.00	A
Grade 6: White/Hispanic Comparisons										
Focus	3.47	1	0.06	3897	1308	0.03	0.01	2.04	0.04	A
Support	10.31	1	0.00	3897	1308	0.03	0.01	3.18	0.00	A
Organization	1.61	1	0.20	3897	1308	0.01	0.01	1.61	0.11	A
Conventions	32.09	1	0.00	3897	1308	-0.07	0.01	-5.40	0.00	A
Integration	2.57	1	0.11	3897	1308	-0.04	0.01	-3.22	0.00	A
Grade 8: Male/Female Comparisons										
Focus	0.10	1	0.76	3767	3571	0.00	0.01	0.54	0.59	A
Support	0.85	1	0.36	3767	3571	-0.01	0.01	-1.42	0.15	A
Organization	0.32	1	0.57	3767	3571	0.00	0.01	0.45	0.65	A
Conventions	0.19	1	0.66	3767	3571	0.00	0.01	-0.14	0.89	A
Integration	2.60	1	0.11	3767	3571	-0.01	0.01	-1.03	0.30	A
Grade 8: White/Black Comparisons										
Focus	0.00	1	0.97	4199	1470	0.00	0.01	-0.09	0.93	A
Support	56.82	1	0.00	4199	1470	0.07	0.01	7.18	0.00	A
Organization	2.08	1	0.15	4199	1470	0.01	0.01	1.23	0.22	A
Conventions	90.19	1	0.00	4199	1470	-0.12	0.01	-9.14	0.00	A
Integration	37.73	1	0.00	4199	1470	-0.02	0.01	-2.17	0.03	A
Grade 8: White/Hispanic Comparisons										
Focus	4.38	1	0.04	4199	1273	0.02	0.01	2.00	0.04	A
Support	43.22	1	0.00	4199	1273	0.06	0.01	6.47	0.00	A
Organization	0.15	1	0.69	4199	1273	0.00	0.01	0.20	0.84	A
Conventions	97.43	1	0.00	4199	1273	-0.12	0.01	-9.39	0.00	A
Integration	22.19	1	0.00	4199	1273	-0.03	0.01	-3.04	0.00	A

2. RELIABILITY and GENERALIZABILITY

The reliability of a test reflects the degree to which test scores are free from errors of measurement that arise from various sources. Test reliability indicates the extent to which differences in test scores reflect real differences in the construct being measured across some variation in one or more factors such as time or specific test items used. Different coefficients can be distinguished accordingly. For example, test-retest reliability measures the extent to which scores remain constant over time. A low test-retest reliability coefficient means that a person's scores are likely to shift unpredictably from one time to another. Generalizability, which may be thought of as a liberalization of classical theory (Feldt & Brennan, 1989, p. 128), treats these error components and their impact on score precision singly and in interaction.

Internal Consistency of Overall Scores

Because achievement test items typically represent only a relatively small sample from a much larger domain of suitable questions, the consistency or generalizability of test scores across items is of particular interest. That is, how precisely will tests line up students if different sets of items from the same domain are used? Unless the lineups are very similar, it is difficult or impossible to make educationally sound decisions on the basis of test scores. This characteristic of test scores is most commonly referred to as *internal consistency*, which is quantified in terms of an index called coefficient alpha. The coefficient, which can range from 0.00 to 1.00, corresponds to a generalizability coefficient for a person by item design or more broadly as a generalizability coefficient for the person by Item by occasions design with one fixed occasion and *k* randomly selected items (Feldt & Brennan, 1989, p 135). Table 2.1 presents alpha coefficients. As the table shows, ISAT tests are highly reliable, since the alpha coefficients are comparable or higher than those typically reported in the literature.

Table 2.1
Reliability Estimates

Grade	Writing	N
05	.90	138,785
06	.90	146,767
08	.90	152,137

Writing scores are affected by other sources of variance, particularly readers (raters), since different readers evaluate different students and prompts. The effect attributable to prompts is important for students at all grades. However, it can only be evaluated directly for 6th- and 8th-grade students who wrote on two different prompts.

Interrater Agreement. Interrater agreement evaluates the consistency of scores assigned to the same essay by different readers. Interrater agreement was monitored daily, and two readers independently scored 10% of the student

essays across grades and prompts. The interrater agreement coefficients for all features and discourse modes are summarized in Table 2.2. The results for the interrater agreement on double-scored papers exceeded the minimum acceptable level of agreement (90% agreement within one point). Scores across raters agree within one point at least 94% of the time.

Table 2.2
Interrater Agreement for Writing Scores

Grade/Discourse Mode	Score	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 5/Expository	Focus	75	22	98
	Support	70	30	100
(N = 28,014)	Organization	69	31	100
	Conventions	73	27	100
	Integration	70	30	100
Grade6/Narrative	Focus	57	37	95
	Support	60	37	96
(N =29,854)	Organization	57	39	99
	Conventions	63	36	99
	Integration	59	38	96
Grade6/Persuasive	Focus	71	23	94
	Support	58	39	97
(N =29,854)	Organization	59	38	97
	Conventions	62	38	100
	Integration	60	38	98
Grade 8/Narrative	Focus	61	35	97
	Support	66	32	99
(N = 30,824)	Organization	62	35	97
	Conventions	67	32	100
	Integration	66	33	98
Grade 8/Persuasive	Focus	62	35	97
	Support	62	37	98
(N = 30,824)	Organization	62	36	98
	Conventions	72	35	99
	Integration	64	35	99

In addition to agreement across raters, writing scores are checked against a standard, or "validation," set of papers. Scoring Directors scored papers and assembled validation sets by closely following the scoring guidelines established by the Validation Committee. Essay packets, each containing 10 essays, were circulated among the readers. Essays for these check sets were chosen to represent a range of score points in all categories.

Readers encountered the validation packets at random intervals throughout the scoring, and some encountered several packets during the scoring process. Readers were unaware of the scores assigned to the papers by the committee. The extent of agreement between a reader's scores and the scores assigned to the papers was calculated every day during the scoring and shared with the readers. This process allowed for the monitoring of reader scoring. The results for all grades, features, and discourse modes are summarized in Table 2.3. Again, the results exceeded the minimum acceptable level of agreement (90% agreement within one point). The agreement of readers with validation papers was higher than the interrater agreement. This is possibly attributable to the

fact that the validation papers are specifically selected to illustrate all points on the scoring scale. The papers that are selected for double scoring, on the other hand, represent a more nearly random selection of papers and scores. Consequently, they are likely to include proportionately fewer extreme scores (e.g., 1, 6), on which there is likely to be higher agreement between raters.

Table 2.3
Agreement with Validation Papers for Writing Scores

Grade/Discourse Mode	Score	% Exact Agreement	% Adjacent Agreement	% Exact + Adjacent
Grade 5/Expository	Focus	85	14	100
	Support	79	21	100
(N = 3,880)	Organization	81	19	100
	Conventions	84	16	100
	Integration	81	19	100
Grade 6/Narrative	Focus	66	31	97
	Support	72	27	99
(N =2,910)	Organization	69	28	98
	Conventions	73	26	100
	Integration	72	26	99
Grade 6/Persuasive	Focus	79	18	97
	Support	71	28	99
(N = 2,850)	Organization	72	26	99
	Conventions	73	27	100
	Integration	73	26	99
Grade8/Narrative	Focus	79	20	99
	Support	89	11	100
(N =3,000)	Organization	85	15	99
	Conventions	78	22	100
	Integration	89	11	100
Grade 8/Persuasive	Focus	75	25	100
	Support	77	22	99
(N = 3,060)	Organization	77	22	99
	Conventions	77	23	100
	Integration	79	21	100

Reliability of the Performance Category Decisions

Students' ISAT scores are reported relative to four performance categories: Academic Warning, Below Standards, Meets Standards, and Exceeds Standards. Sets of score cutoffs were developed for each learning area and each grade. The development of the score cutoffs that define these categories is fully documented in separate publications available from ISBE (*Performance Levels for the Illinois Standards Achievement Tests: Reading, Mathematics, Writing* and *Performance Levels for the Illinois Standards Achievement Tests: Science, Social Science*). However, the process may be briefly described as follows.

Prior to the meetings of the standard-setting panels themselves, which took place during April 1999, ISBE convened committees of curriculum experts to develop concrete descriptions of student knowledge and skill levels that define the specific performance categories (achievement descriptors). Educators throughout Illinois extensively reviewed these descriptions.

Panels of recognized subject matter experts convened in Springfield to translate the verbal descriptions into cut scores on the ISAT tests (i.e., scores that define the boundaries between categories). Panelists were drawn from a pool of educators who had specific knowledge of student performance at the grade levels being assessed by ISAT and experience in assessing students at those grade levels. Panelists were selected to be broadly representative of the geographic and ethnic diversity of Illinois' public school system. A total of 62 writing educators participated in the standard-setting process.

The panelists worked iteratively and evaluated both the rating scales and student writing samples from the perspective of the achievement descriptors. Item performance statistics were provided to help panelists anchor their ratings. The cutoff scores that resulted are shown in Table 2.4. Results of applying these cutoffs to the 2008 test population are shown later.

The reliabilities of such classifications, which are criterion-referenced, are related to the reliabilities of the tests on which they are based, but they are not equivalent to the test reliabilities, which are based on norm-referenced measurement. Glaser (1963) was among the first to draw attention to this distinction, and Feldt and Brennan (1989) extensively reviewed the topic.

Table 2.4
ISAT Cutoffs for Each Performance Level

Grade	Academic Warning	Below Standards	Meets Standards	Exceeds Standards
05	6-13	14-20	21-27	28-33
06	6-13	14-20	21-27	28-33
08	6-14	15-20	21-27	28-33

As Feldt and Brennan (1989, p. 140) point out, approaches to the development of reliability coefficients for criterion-referenced interpretations of test scores have been based either on squared-error loss or threshold loss. It is threshold loss, which evaluates the consistency with which people are consistently classified with respect to a criterion, that is of greater concern here. Specifically, the issue is how consistently do tests classify students with respect to the performance standards?

Two threshold-loss coefficients have been developed: p, the proportion of persons consistently classified on two parallel tests, and k (kappa), which corrects p for the proportion of consistent classifications that would be expected by chance. Because scores on classically parallel tests are rarely available in practice, methods have been developed to estimate these values from a single test (Subkoviak, 1984). An approach proposed by Peng and Subkoviak (1980) was applied to the performance classifications made on the basis of the 2008 tests.

Table 2.5 presents these values for p, k, and p_{miss}, the expected proportion of inconsistent decisions, which is simply (1 - p). In interpreting the first two indexes, Feldt and Brennan (1989) suggest that p reflects the *consistency of*

decisions made about examinees, whereas *k*, since it is corrected for chance, reflects the *contribution of the test* to the consistency of the decision.

Overall, the values suggest that decisions made with respect to the student performance classifications would be very consistent. Note that the *p* and *k* values are calculated for the complete test population. Values for other test populations (e.g., IEP students, ELL students) may differ.

Table 2.5
Reliability of Student Performance Decisions

		Academic Warning/Below Standards			Below Standards/Meets Standards			Meets Standards/Exceeds Standards		
Area	Grade	P	kappa	p_{miss}	p	kappa	p_{miss}	p	kappa	p_{miss}
Writing	5	0.954	0.632	0.046	0.862	0.713	0.138	0.914	0.678	0.086
	6	0.964	0.614	0.036	0.862	0.713	0.138	0.952	0.650	0.048
	8	0.970	0.609	0.030	0.866	0.710	0.134	0.936	0.662	0.064
AVERAGE		0.963	0.618	0.037	0.863	0.712	0.137	0.934	0.664	0.066

3. VALIDITY

Test validity refers to the degree that a test measures what it is intended to measure. Evidence that supports a test's validity is gathered for different aspects and through different methods. The three most recognized aspects are content validity, construct validity, and criterion-related validity. Content validity refers to how well a test covers the content of interest. The process does not involve any statistical computation. Instead, it examines the correspondence between test blueprints that describe the intended content and test items. Construct validity is comprised of the analyses of a test's internal constructs in order to confirm that the test indeed functions as it is intended to function. Analyses of construct validity include correlations between items and the test, discrimination between subgroups, factor analysis, and multitrait-multimethod methods. Criterion-related validity indicates whether a test is consistent with other tests that measure the same content. Depending on the use of information, criterion-related validity can be either concurrent or predictive. The former focuses on the relationship between two tests given at the same time that measure the same content and the later focuses on using a test to predict future performance (Cronbach & Meehl, 1955; Crocker & Algina, 1986; and Clark & Watson, 1995).

Content Validity

Evidence of content validity have been provided in the 2006 Test Construction Specifications, which contains descriptions of the blueprint, the process, and the decisions made for defining and developing the test.

Construct Validity

DIMENSIONALITY

Dimensionality is a unique aspect of construct validity. Achievement tests are usually intended to measure a unidimensional construct. Although it is generally agreed that unidimensionality is a matter of degree rather than an absolute situation, there is no consensus on what defines dimensionality or on how to evaluate it. Approaches that evaluate dimensionality can be categorized into answer patterns, reliability, components and factor analysis, and latent traits. Component analysis and factor analysis is the most popular method for evaluation (Hattie, 1985; Abedi, 1997).

Lord (1980) stated that if the ratio of the first to the second eigenvalue is large and the second eigenvalue is close to other eigenvalues, the test is unidimensional. Divgi (1980) expanded Lord's idea and created an index by considering the pattern of the first three factor components (eigenvalues). The Divgi Index examines the ratio of the difference of the first and second eigenvalues over the difference of the second and third eigenvalues. A large ratio indicates a greater difference between the first and second eigenvalues, thus,

creating a unidimensional tendency. A cut value of 3 is as the minimum criterion for declaring unidimensionality.

The results reported in Table 3.1 resulted from a principal axis factoring estimation procedure of the unreduced correlation matrices. The values for all three tests is greater than 3. The Scree plots, which are shown in Figures 3.1 through 3.3, similarly reveal a large first component and relatively trivial secondary components, thereby supporting a unidimensional test structure.

Table 3.1
Divgi Index

Grade	Index
5	14.8
6	10.5
8	9.5

Figure 3.1
Grade 5 Scree Plot

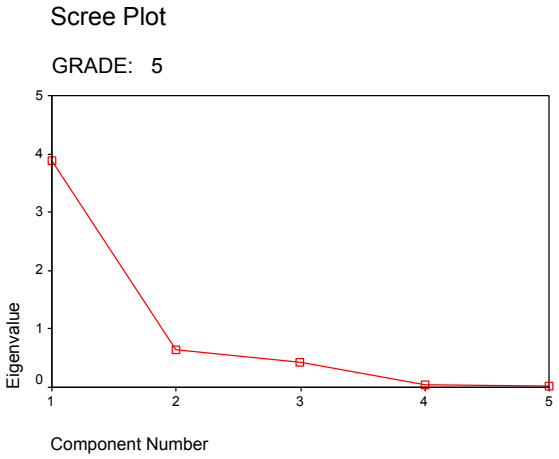


Figure 3.2
Grade 6 Scree Plot

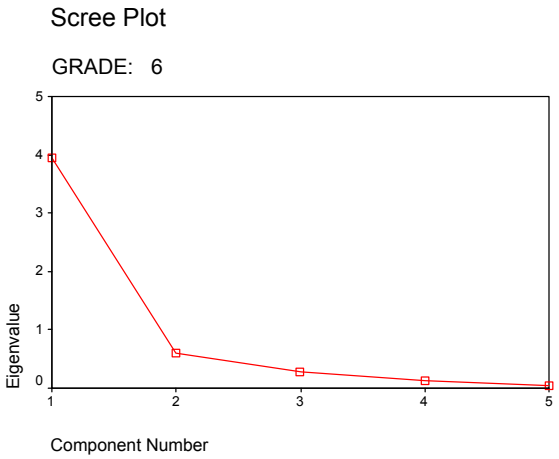
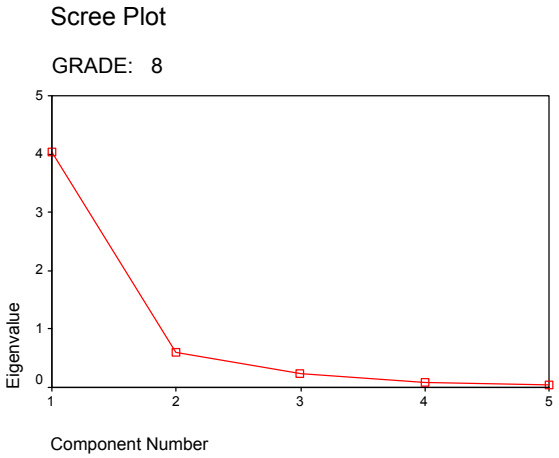


Figure 3.3
Grade 8 Scree Plot



INTERNAL CONSTRUCT

The purpose of studying the internal structure of a test is to demonstrate that all of the elements work coherently. Methods that are used to provide evidence of the internal structure of a test are usually associated with correlations, for example, the item-total correlation and subscale-total correlation.

Empirical data is used to evaluate test structure through point-biserial correlations of item-total and subscale-total correlations. The subscale scores are the points earned for each reporting category. The corrected point-biserial, in contrast to the uncorrected method, excludes an item from the total score when computing its point-biserial. This method avoids the overestimation issue that commonly occurs in the uncorrected method. The subscale-total correlation includes the subscale items in the total scores. A summary of item-

total point-biserial correlation by grade is listed in Table 3.2. The strong correlations indicate that individual items as well as subscales work coherently.

Table 3.2
Item-Total Point-Biserial Correlation Coefficients

	Grade 5	Grade 5	Grade 8
Focus	.7035	.8053	.8375
Support	.9492	.9168	.9280
Organization	.9511	.9232	.9460
Conventions	.5542	.5748	.5737
Integration	.9376	.9352	.9558
Median	.9376	.9168	.9280

4. RESULTS

Table 4.1 shows results for writing with respect to the performance standards. Table 4.2 summarizes results with respect to writing feature scores. Note that Conventions is scored on a three-point scale while all other features are scored on a six-point scale.

Table 4.1
Percentages of Students by Grade Falling into Each Performance Level

Grade	Academic Warning	Below Standards	Meets Standards	Exceeds Standards	Meets + Exceeds Standards
Grade 5					
2007	9.6	40.3	39.2	10.9	50.1
2008	9.0	35.7	45.3	10.0	55.3
Grade 6					
2008	5.4	34.7	55.2	4.7	59.9
Grade 8					
2007	5.7	30.7	54.1	9.5	63.6
2008	5.7	31.0	54.3	8.9	63.2

Note: Because of rounding, the percentages in each row may not total exactly to 100%.

Table 4.2
Mean Writing Feature Scores of Students by Prompt

Grade	Type	F	S	O	C	I
05	E	4.7	3.6	3.5	2.5	3.6
06	N	4.6	3.5	3.5	2.4	3.6
06	P	3.7	3.6	3.5	2.3	3.6
08	N	3.9	3.6	3.6	2.6	3.6
08	P	3.9	3.9	3.9	2.5	3.9

Note: Prompt type: P = Persuasive; E = Expository; N = Narrative

REFERENCES

Abedi, J. (1997). *Dimensionality of NAEP Subscale Scores in Mathematics*. CSE Technical Report 428. http://www.cse.ucla.edu/CRESST/pages/reports.htm.

Clark, L. A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7 (3), 309-319.

Crocker, L. M. & Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. Orlando, FL: Harcourt Brace Jovanovich, Inc.

Cronbach, L. J. & Meehl, P. E. (1955). Classics in the History of Psychology. http://psychclassics.yorku.ca/cronbach/construct.htm.

Divgi, D. R. (1980). Dimensionality of Binary Items: Use of a Mixed Model. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement (3rd Edition)* (pp. 105-146). New York: Macmillan.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9 (2), 139-164.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. New York: Erlbaum Associates.

Peng, C-Y, J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

Subkoviak, M. J. (1984). Estimating the reliability of mastery/non-mastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267-291). Baltimore: Johns Hopkins Press.