DOCUMENT RESUME

ED 343 933 TM 018 085

AUTHOR Linn, Robert L.; And Others
TITLE Cross-State Comparability of Judgments of Student Writing: Reports from the New Standards Project.
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-335
PUB DATE 91
CONTRACT R117G10027
NOTE 54p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; Correlation; Educational Assessment; Elementary Secondary Education; Evaluators; *National Competency Tests; *Scoring; *State Programs; State Standards; Student Evaluation; Test Construction; *Testing Programs; Test Results; *Writing Evaluation
IDENTIFIERS *New Standards Project (LRDC); *Performance Based Evaluation

ABSTRACT

The New Standards Project is a joint effort of the Learning Research and Development Center (LRDC) at the University of Pittsburgh (Pennsylvania) and the National Center on Education and the Economy toward creation of a national examination system based on performance assessments. This study explored the feasibility of comparing performance on different tasks across states using writing assessments, the current assessments closest to the type of assessment that the Project seeks to encourage. Ten states (Arizona, California, Colorado, Connecticut, Maine, New York, Oregon, South Carolina, Texas, and Vermont) participated in a cross-state scoring workshop in July 1991. Writing projects from elementary school, middle school, and high school levels were submitted, using selections from each state's regular operational writing assessment program. Each state also sent experienced readers. Eight sets of papers from elementary schools, eight sets from middle schools, and seven sets from high schools were multiply scored. Correlations of scores assigned by readers from one state with those from another state were generally quite high. The high level of agreement provides good documentation of the fact that states share a common view of the relative ordering of students' writing from low to high. It is evident, however, that considerable cross-state discussion would be required to arrive at common performance standards. Three figures and 21 tables illustrate the discussion, and a 10-item list of references is included. (SLD)

Cross-State Comparability of Judgments of Student Writing: Results from the New Standards Project

CSE Technical Report 335

Robert L. Linn

National Center for Research on Evaluation, Standards and Student Testing (CRESST)
UCLA Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310)206-1532

Acknowledgement[1]

This study would not have been possible without the dedication and hard work of a large number of people. The readers from the ten participating states were an extraordinarily hard-working and dedicated group of professionals. Due to differences in the nature of the writing assignments to be scored from those in their own state assessment programs, the limitations of time, and questions about the nature of the overall exercise, the readers' task was quite demanding and sometimes frustrating. Nonetheless, the quality of the data that resulted from their work is a tribute to the seriousness of their efforts. Support of the staff from the New Standards Project was excellent and essential to the successes of the effort. We particularly wish to thank Eric Hamilton, Sally Jacob, and Natalie Peterson in this regard. Finally, we thank the leadership of the New Standards Project, Dan Resnick, Lauren Resnick, and Mark Tucker, for making the workshop a reality.

Robert L. Linn, Project Director
University of Colorado at Boulder

and Co-Authors:

Vonda L. Kiplinger
University of Colorado at Boulder

Carmen W. Chapman
Illinois State Board of Education

Paul G. LeMahieu
Pittsburgh Public Schools

---

[1] The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education. It was also conducted with partial support from the John D. and Catherine T. MacArthur Foundation and the Pew Charitable Trusts through the New Standards Project. The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education or of other aforementioned funding agencies.

Cross-State Comparability of Judgments of Student Writing: Results from the New Standards Project Workshop

Robert L. Linn

Executive Summary

The New Standards Project is a joint effort of the Learning Research and Development Center at the University of Pittsburgh and the National Center on Education and the Economy. The project is an effort to create a national examination system that is intended to serve as a catalyst for major educational reform. One of several unique features of the proposed national examination system is its reliance on performance tasks that are determined individually by local districts, states or clusters of states so that the examinations can be closely tied to the curriculum in use in the relevant educational setting. Nonetheless, the performance of students within any participating partner would be compared to a common national performance standard.

The purpose of this study is to explore the feasibility of comparing performance on different tasks across states. Writing assessment is the one area where current assessments in a number of states are closest to the type of performance-based examination tasks that the New Standards Project seeks to encourage. Hence, writing assessment was a natural choice for an initial exploration of the degree to which there is sufficient cross-state comparability to support the idea of developing a common national standard against which assessments from different settings could be compared.

Ten states (Arizona, California, Colorado, Connecticut, Maine, New York, Oregon, South Carolina, Texas, and Vermont) participated in a cross-state scoring workshop in July 1991. Writing products from three separate grade levels—elementary, middle school, and high school—were used in the workshop and states participated at one or more of these three grade levels. Participating states provided a sample of student writing products produced in response to a single prompt or writing assignment as part of the state's regular operational writing assessment program. A list of paper numbers

and the operational scores assigned to those papers also was provided by participating states.

At each grade level where a participating state provided a set of student papers, the state also sent a delegation of experienced readers of the state's writing assessment to the workshop. Three cross-state scoring sessions were held during the course of the four-day workshop. At each scoring session, the readers from a given state received a set of papers from another state. Those papers were read and scored using the standard procedures of the scoring state. At the completion of the workshop the papers from each providing state had been independently scored by readers from three other states. Those three sets of scores were compared to each other and to the operational scores that previously had been obtained from each state providing papers for the workshop.

In all, eight sets of papers written by students in the elementary grades (grades 3 to 5) were each scored by readers from four different states. At the middle school level (grades 7 and 8) there also were eight sets of papers that were multiply scored, while at the high school level there were seven sets of multiply scored papers.

Correlations of the scores assigned by readers from one state with those assigned by readers from another state were generally quite high. The median of 48 cross-state correlations for the eight elementary level paper sets was .73. The corresponding medians for 45 middle school and 42 high school cross-state correlations were .80 and .81, respectively. These high cross-state correlations indicate that a substantial consensus already exists regarding the relative order of low to high quality of student writing at all three grade levels. The high correlations between the scores assigned by readers from different states were achieved without modification of the scoring procedures used by the 10 states that participated in the workshop. Nor was there any attempt to select writing prompts or assignments that were most compatible with those used by other states. In view of the substantial differences across states in the nature of the writing assignments, the uses that are made of the scores, the scoring procedures, and the age of the state writing assessment programs, the high level of agreement that was achieved provides good documentation of the fact that these states share a common view regarding the relative ordering of student writing from low to high. Even though the scales may not all

represent the full range of writing competence, the criteria of poor to good writing is ordered in quite similar ways across the 10 participating states.

The high level of agreement on the r lative ordering of papers by readers from different states that was achieved in this study is a prerequisite for any system that would compare results within a state to a common national standard. *But the agreement on relative order of papers is not the same as agreeing on absolute judgments in comparison to a set standard of mastery.* There were some rather large differences in the level of scores that were assigned to papers even in instances where the relative ordering of papers by different states was almost as similar as the within-state rating reliabilities would allow. Even states that use the same number of score points, most commonly 1 to 4, sometimes differed substantially in the implicit standard with which the readers assigned the scores. In extreme cases, it appears be about as difficult for a paper to earn a score of 3 according to the standards of one state as it is to earn a 4 according to the standards of another state.

Such differences in apparent stringency of scoring do not preclude the possibility of agreeing on a common standard. In the above extreme case, for example, it is conceivable that a score of 4 would be required as the standard of "excellent" in the state with the more lenient scoring, whereas a 3 or higher would be required in the state with the more stringent scoring. It seems clear, however, that considerable cross-state discussion (what in some contexts is referred to as "moderation" of scoring) would be required to arrive at common performance standards, that is, agreement on a standard requires specific definitions of what score points represent.

CROSS-STATE COMPARABILITY OF JUDGMENTS OF STUDENT WRITING: RESULTS FROM THE NEW STANDARDS PROJECT

Robert L. Linn

The New Standards Project is a joint effort of the Learning Research and Development Center (LRDC) at the University of Pittsburgh and the National Center on Education and the Economy (NCEE) supported by grants from The MacArthur Foundation and The Pew Charitable Trusts. The major goal of the project is the creation of a national examination system that "will be part of an educational system structured to reward students who work hard to meet a clear mastery standard and provide school staff with the resources and incentives they need to help all students meet that standard" (LRDC & NCEE, no date).

The words "examination" and "system" are intended to distinguish the proposed development from current testing practices and from proposals for a single national test. The examination system that is envisioned by the project's co-directors, Lauren Resnick and Mark Tucker, is quite different from most testing programs that currently exist in this country (LRDC & NCEE, no date; Resnick, 1991; Tucker, 1991).

A critical difference between the described examination system and the tests that are now most widely used is that the proposed examinations would be syllabus driven. Students would be expected to study and practice in preparation for the examinations, and teachers would be expected to help students prepare for the exams. Therefore, both students and teachers must know the criteria upon which evaluations will be based. The model for the teacher would become more like that of a coach than that of an evaluator.

A second distinguishing characteristic of the proposed examination system is that the examinations would be quite different in form than the familiar multiple-choice test. The examination system would include student projects and portfolios of student work, as well as timed, on-demand performance examinations. These "three P's," portfolios, projects, and

performances, would each be expected to require students to solve challenging problems imbued with meaning in and of themselves. Unlike most testing programs in the schools, the examinations would not occur as one-time events. Rather, the combination of performances, projects, and portfolios would be coordinated components of an evaluation system that is closely coupled to learning activities and desired learning outcomes over an extended period of time. All three components would contribute to the total examination system as it is envisioned by Resnick and Tucker.

A third distinguishing characteristic of the proposed examination system is its reliance on tasks determined locally or by clusters of states and districts rather than reliance on a single set of nationally determined tasks that would be administered to all students. A proposed national standards board could provide oversight and could contribute prototypes of tasks as well as technical assistance, but would not impose a particular set of examination questions or definitions of acceptable projects or portfolios. Instead, national standards would depend on a system of review of products from states or clusters of states and districts and, perhaps, some form of national anchor examination that would be administered on a sample basis.

Comparability

The desire to compare results from examinations composed of different tasks and assignments against a common national standard poses a substantial challenge for the proposed system. Test equating is the approach normally used when scores for individuals or groups obtained from different tests need to be compared. The conditions under which equating is considered equitable were defined strictly by Fred Lord as follows. "If an equating of tests x and y is to be considered equitable for each ... [examinee], it must be a matter of indifference to ... [examinees or anyone using the test results] whether they are to take test x or test y" (Lord, 1980, p. 195). Lord went on to demonstrate that fallible tests cannot be equated in the above sense unless the tests are strictly parallel, in which case, no equating is needed. In other words, if you need to equate, you can't, at least not in the strict sense.

In practice, of course, equating techniques are applied to tests that fall short of being parallel. How well the techniques work, however, depends heavily on the degree to which the tests approximate the ideal of being parallel.

Examinations consisting of portfolios of student work, individually tailored projects, and complex performances in response to on-demand tasks that are defined locally or by clusters of states and districts will certainly not yield scores that can be treated as arising from measures that even approximate the criteria for parallel measures. At a minimum, equating techniques assume that the tests measure the same construct. However, the different types of assessments discussed above surely will not measure a single construct. In short, strictly equated scores simply cannot be achieved as a part of the proposed national examination system.

A more realistic goal is to provide a reasonable degree of assurance that student performance satisfies or fails to satisfy some judgmental standard of quality. Comparability of results of examinations composed of different components in terms of the national standards will depend heavily on human judgment. It may be more reasonable, for example, to agree on high standards of excellence for both high jumping and figure skating than to claim that two performances are equivalent. Procedures that are sometimes referred to as calibration, sometimes as moderation, and sometimes as verification can assist the judgmental process and provide checks and balances, but they will not remove the need to rely on human judgment at each stage of the process. In fact, the conception of learning and what stands as valued knowledge also is implicit in any examination system. The role of judgment, however, is more obvious at all stages for the types of assessments envisioned by the New Standards Project. Nonetheless, clearly articulated criteria will be required for raters to judge whether a standard has been achieved.

For such a bottom-up system of examinations to satisfy the desire for national standards, there must be a reasonable consensus that student work on different tasks is sufficiently comparable in terms of the quality of thinking and demonstration of accomplishment to be judged in terms of a common standard. The purpose of this paper is to describe the results of The New Standards Project's initial effort to evaluate the degree to which current judgmental scoring procedures used by different states are comparable.

Workshop on Cross-State Grading of Students' Writing

The prime example of performance examinations already in widespread use is in the area of writing. A number of states and districts as well as

commercial test publishers have introduced direct assessments of writing as a component of their operational testing programs. Consequently, writing assessment comes the closest to having in place, on a wide scale basis, at least one of the "three P's"—timed, on-demand performance.

Because of the extensive experience that some states have had in the direct assessment of writing, writing assessment was a natural starting point for the New Standards Project. Hence writing was the area selected for the initial investigation of the comparability of judgments of student products using the standards and procedures of several different states.

With the support of the New Standards Project, experienced readers of student writing representing 10 states participated in a cross-state scoring workshop in July, 1991. Writing products from three separate grade levels, elementary (grades 3 to 5), middle school (grades 7 or 8), or high school (grades 11 or 12), were used in the workshop, and states participated in one or more of the three levels. The participating states and the grade levels at which they participated are shown in Table 1. (All tables are provided at the end of this document.) As can be seen, a total of eight states participated at each of the two lower grade levels and seven states participated at the high school level.

States participating in the workshop provided a sample set of student writing examination papers that previously had been scored as part of the normal state writing assessment. States were asked to provide 45 exam papers at each of the elementary and middle school grade levels and 36 papers at the high school level. The smaller number of papers was used at the high school level because the average time to score high school level papers is longer than that for the lower grades. The states were instructed to select papers to represent an even distribution of all score levels used by the state. The papers provided by a state could be responses to narrative, descriptive, expository, or persuasive writing tasks, but all papers for a given grade level were to be written in response to a single prompt or writing assignment.

States provided a list of paper identification numbers and the operational scores assigned to each paper. Scores or other comments were not shown on the papers, however. In addition to sets of papers and scores, states were asked to provide descriptions of their writing assessment program, their scoring rubrics, and materials used to train readers to use their scoring procedures.

Table 1

States Participating in the Cross-State Scoring Workshop

State	Elementary (Grades 3-5)	Middle School (Grades 7-8)	High School (Grades 11-12)
Arizona	X	X	X
California		X	X
Colorado	X		
Connecticut	X	X	
Maine	X	X	X
New York	X		X
Oregon	X	X	X
South Carolina		X	X
Texas	X	X	X
Vermont	X	X	
Total by Level	8	8	7

States were asked to send delegations to the workshop consisting of a state-level writing curriculum supervisor, a state-level evaluation expert, and current classroom teachers who were experienced readers of the state writing examinations. Each state was asked to include in its delegation three experienced readers for each grade level at which the state was participating, although not all states were able to provide a complete delegation.

Scoring systems. As can be seen in Table 2, the 10 participating states use a variety of scoring systems. Although the assignment of a single holistic score is the most common procedure among the participating states, separate scores for Content and Mechanics are assigned by Arizona and for Rhetorical Effectiveness and Conventions by California. Vermont and Oregon, on the other hand, score papers on five and seven analytical dimensions, respectively.

States differ in terms of the number of scale points that are used (1 to 4, 1 to 5, 1 to 6, and 0 to 100). Indeed, Vermont does not use numbers at all. Instead, papers are assigned a rating of "rarely," "sometimes," "frequently," or "extensively" on each of the five dimensions. The resolution procedures for discrepant scores assigned by two independent readers also vary from state to state. Maine, for example, uses the sum of scores assigned on a 1 to 6 scale by two independent readers, which yields a final score ranging from 2 to 12. In Oregon and Vermont scores assigned by both raters are reported separately if they are within one scale point of each other and a resolution score is assigned only for cases where the scores differ by more than one score category. In other cases, the third reader determined the resolution score.

Cross-state scoring. Three cross-state scoring sessions, of approximately three hours each, were held during the course of the four-day workshop. At each scoring session, the readers from a given state received a set of essays from another state. The complete design for distribution of papers from states providing papers to groups of readers from three other states is shown in Table 3. As can be seen in Table 3, for example, the elementary level readers from Arizona received the set of papers provided by Colorado in the first scoring session, the set of papers provided by Connecticut in the second scoring session, and the set of papers provided by New York in the third scoring session. In turn, the papers provided by Arizona were scored by

Table 2

Scoring Procedures Used by States Participating in the Cross-State Scoring Workshop

State	Grades	Scores	Scale	Final Scores
Arizona	3, 8, 12	(1) Content (2) Mechanics	1 to 4	Sum of scores on two features
California	8, 12	(1) Rhetorical effectiveness (2) Conventions	1 to 6	Dimensions separately reported
Colorado	4	Primary trait	1 to 4	Two readers, 3rd to resolve discrepancies
Connecticut	4, 8	Holistic	1 to 4	Sum of scores of two readers
Maine	4, 8, 11	Holistic	1 to 6	Sum of scores of two readers
New York	5	Holistic	1 to 4	Sum of scores of two readers
	11	Holistic	0 to 100	Sum of scores on three essays
Oregon	5, 8, 11	(1) Ideas (2) Organization (3) Voice (4) Word Choice (5) Sent. Fluency (6) Conventions (7) Content	1 to 5	Separate ratings reported
South Carolina	8	Holistic	1 to 4	Additional scores for students below a score of 3
Texas	5, 7, 11	Holistic	1 to 4	Based on four criteria
Vermont	4, 8	(1) Purpose (2) Organization (3) Details (4) Voice (5) Grammar	Four verbal categories	Separate ratings reported

Table 3

Design for Rescoring by Grade Level

Elementary Level (Grades 3-5)

State Scoring Papers	State Providing Papers								Number of States Scored
	AZ	CO	CT	ME	NY	OR	TX	VT	
AZ	–	1	2		3				3
CO	1	–		2		3			3
CT			–	1	2		3		3
ME			1	–		2		3	3
NY		3			–	1	2		3
OR	3				1	–		2	3
TX		2		3			–	1	3
VT	2		3				1	–	3

Middle School Level (Grades 7-8)

State Scoring Papers	State Providing Papers								Number of States Scored
	AZ	CA	CT	ME	OR	SC	TX	VT	
AZ	–	1	2		3				3
CA	1	–		2			3		3
CT			–	1	2			3	3
ME			1	–		3	2		3
OR		3			–		1	2	3
SC	2		3			–		1	3
TX	3				1	2	–		3
VT		2		3		1		–	3

High School Level (Grades 11-12)

State Scoring Papers	State Providing Papers							Number of States Scored
	AZ	CA	ME	NY	OR	SC	TX	
AZ	–	1	2	3				3
CA		–	1		2	3		3
ME			–	1	3		2	3
NY		3		–		2	1	3
OR	1			2	–		3	3
SC	2		3		1	–		3
TX	3	2				1	–	3

Note. Numbers indicate the scoring session at which the rating state rated the papers from a given providing state.

readers from Colorado, Vermont, and Oregon in scoring sessions 1, 2, and 3, respectively.

By the end of the workshop, the papers provided by each state had been scored by teachers from three other states. Thus, with the addition of the operational scores that were assigned to the papers prior to the workshop, each set of papers was assigned scores according to the procedures used in a total of four states. One exception to this pattern occurred at the middle school level where operational scores were not provided by Arizona.

In order to achieve the scores that most closely replicated what would be assigned by actual operational ratings by the scoring states, the participants were instructed to duplicate an abbreviated state scoring process within the time and physical constraints of the workshop. Readers of the receiving state were directed to review the prompt from the sending state, classify it within their own scoring rubric for the grade level, and then peruse the papers looking quickly for low, medium, and high papers to score as a group to establish them as anchors or benchmarks. The purpose of this was to make sure all readers were assigning their state scores on an unfamiliar prompt in the most accurate way possible (particularly for prompts that, in some cases, were very different from their own). This "mini-calibration" session was implemented to ensure the highest within-state scoring agreement with the limited amount of time.

Where possible, the remaining papers were scored independently by two readers; a third reader served as the referee when the two independently assigned scores did not agree. This was accomplished by dividing the set of papers from the providing state into three packets of equal numbers of papers. During the first part of a scoring session, reader 1 read papers in the first packet, reader 2 read those in the second packet, and reader 3 read the papers in the third packet. Scores were recorded on separate scoring sheets. The packets of papers were then exchanged (packet 3 to reader 1, packet 1 to reader 2, and packet 2 to reader 3), and the second set of independent scores was obtained. The papers were exchanged a final time and each reader served as the referee for the packet of papers that had been scored by the other two readers.

Throughout the workshop, it was stressed that readers were to apply their own state's scoring procedures as they normally would within their own state assessment. An explicit effort was made to avoid modifying a state's scoring procedure to make it more compatible with that of the state that provided the papers.

Analyses. The primary focus of the analyses was on the refereed scores provided by each state. In cases where the state's referee procedure is simply to record both scores when the two independent readers are within one point of each other, the scores of the two independent readers were summed. This resulted in refereed scores from Oregon readers ranging from 2 to 10, for example. For purposes of the analyses, the Vermont verbal category ratings were also converted to numerical scores (rarely=1, sometimes=2, frequently=3, and extensively=4). As in the case of Oregon, the refereed scores were a sum of the two readings in Vermont so that the final scores ranged from 2 to 8.

For each set of papers, correlations were computed among the operational scores obtained from the state providing the paper and all three sets of refereed scores obtained during the three scoring sessions held during the workshop. Contingency tables also were obtained and used to provide a means of evaluating the degree to which the standards used by different states were comparable. Finally, since the scoring systems used by different states vary in terms of number of scores assigned and the number of scale points used, scores were transformed to a common numerical scale with the minimum score equal to zero and the maximum score equal to one so that means could be compared as an alternative way of judging the relative stringency of the scores assigned by readers from different states.

Results

Cross-State Relationships

States with multiple scores. The multiple scores (reflecting different dimensions assessed) assigned by some states were maintained in the preliminary analyses, but for some of the later cross-state comparisons, a single score was used for each state. The multiple score results at the elementary grade level are presented here in some detail. Multiple score results at the two higher grade levels are presented in less detail except when

the multiple scores reveal different patterns or suggest conclusions that differ from those based on the analyses of the elementary level data.

Three of the states using multiple scores (Arizona, Oregon, and Vermont) participated at the elementary level. The intercorrelations among the multiple scores for any one of these states considered alone can be computed using four sets of papers (the states' own operational scores and the three sets of papers that readers from that state scored).

Correlations between the Arizona Content and Mechanics scores were .39, .55, .63, and .78 for the four different sets of papers. Due to missing scores for one data set, the .39 was based on only 22 observations, while the other three were based on either 44 or 45 observations. While apparently giving distinct information, the two Arizona scores are nonetheless substantially related.

The intercorrelations among the seven refereed scores for the four elementary level sets of papers scored by the Oregon readers are listed in Table 4. Also shown in Table 4 are the mean values of the four correlations between each pair of scores using Fisher's Z transformation. From an inspection of Table 4, it can be seen that the correlations among the pairs of scores in Oregon are generally quite high. The variables have been arranged in Table 4 in an order to make the pattern of highest correlations more apparent. The first four scores listed in Table 4 (Ideas, Organization, Content and Voice) have average intercorrelations across the four sets of papers for all pairs of variables that are between .84 and .88. The remaining three scores (Conventions, Fluency and Word Choice) have somewhat lower correlations with the first four variables (averages between .71 and .81) and with each other (averages between .69 and .77).

Table 5 displays a comparable set of intercorrelations among the five Vermont scores. The pattern of correlations for the Vermont readers is similar to that of the Oregon raters in that the first four scores (Purpose, Organization, Details and Voice), with one exception (Organization with Voice), all have average intercorrelations with each other that are .8 or higher. The Grammar scores have lower correlations with the first four scores than those scores have with each other, however.

The patterns of correlations in Tables 4 and 5 suggest that there may be one major overall quality dimension (measured primarily by the first four

Table 4

Intercorrelations Among the Oregon Elementary Scores for Four Sets of Papers (Below the Diagonal) and the Mean Correlations Based on Fisher's Z Transformations (Above the Diagonal)

		Score						
Score	Data Set	Ideas	Organ	Cont	Voice	Conv	Fluen	Word
Ideas	AZ	–						
	NY	–	.88	.86	.84	.72	.81	.74
	OR	–						
	VT	–						
Organi-	AZ	.87	–					
zation	NY	.84	–	.87	.84	.70	.79	.74
	OR	.92	–					
	VT	.89	–					
Content	AZ	.91	.87	–				
	NY	.72	.80	–	.86	.71	.74	.75
	OR	.91	.94	–				
	VT	.85	.84	–				
Voice	AZ	.88	.85	.87	–			
	NY	.71	.72	.77	–	.71	.75	.75
	OR	.89	.92	.91	–			
	VT	.84	.81	.84	–			
Conven-	AZ	.67	.63	.61	.66	–		
tions	NY	.65	.72	.73	.75	–	.77	.69
	OR	.78	.74	.76	.73	–		
	VT	.77	.70	.74	.69	–		
Fluency	AZ	.71	.69	.63	.65	.72	–	
	NY	.73	.84	.76	.76	.72	–	.75
	OR	.86	.81	.82	.81	.86	–	
	VT	.88	.80	.74	.75	.76	–	
Word	AZ	.58	.63	.66	.56	.55	.73	–
Choice	NY	.67	.68	.56	.76	.64	.71	–
	OR	.86	.86	.84	.87	.82	.82	–
	VT	.78	.73	.76	.73	.71	.74	–

Note. Organ = Organization; Cont = Content; Conv = Conventions; Fluen = Fluency; Word = Word Choice.

Table 5

Intercorrelations Among the Vermont Elementary Scores for Four Sets of Papers (Below the Diagonal) and the Mean Correlations Based on Fisher's Z Transformations (Above the Diagonal)

		Score				
Score	Data Set	Purpose	Organ	Details	Voice	Grammar
Purpose	AZ	–				
	CT	–	.87	.86	.82	.73
	TX	–				
	VT	–				
Organi-	AZ	.91	–			
zation	CT	.85	–	.81	.75	.72
	TX	.80	–			
	VT	.89	–			
Details	AZ	.87	.80	–		
	CT	.77	.82	–	.82	.71
	TX	.81	.75	–		
	VT	.94	.87	–		
Voice	AZ	.71	.63	.73	–	
	CT	.81	.85	.85	–	.69
	TX	.81	.62	.81	–	
	VT	.91	.83	.88	–	
Grammar	AZ	.60	.69	.61	.49	–
	CT	.68	.67	.67	.63	–
	TX	.67	.65	.65	.62	–
	VT	.88	.84	.84	.89	–

Note. Organ = Organization.

scores as arranged in Tables 4 and 5) and one or more additional dimensions concerned with conventions, grammar, or other specific characteristics. We do *not* mean to suggest by this observation that either Oregon or Vermont should necessarily alter their scoring to produce fewer scores. The multiple scores may have utility for instructional purposes that we have not considered. For our focus on cross-state comparisons, however, not all scores are needed. Moreover, the pattern of consistently high intercorrelations suggests that the number of scores can be reduced without excessive modification of the evaluative (if not the instructional) structures of the assessment scheme.

Patterns of correlations among the seven Oregon scores are generally similar at the higher grade levels to the ones reported in Table 4. For example, the average correlations among Ideas, Organization, and Content over four paper sets per grade level ranged between .83 and .87 at the middle school level and between .81 and .86 at the high school level. There are two exceptions to the otherwise similar patterns of correlations, however. First, the Voice scores have higher correlations with the scores for Ideas, Organization, and Content in Table 4 than those at the middle school level. For the four middle school sets of papers, the average correlations of Voice with Ideas, Organization, and Content are .74, .66, and .67, respectively. These averages are a good deal lower than their counterparts at the elementary level, where they averaged .84 to .86. For the four high school level paper sets, the average correlations of Voice with Ideas, Organization, and Content fall between those at the elementary and middle school levels (.81, .76, and .73, respectively).

The second exception concerns the Fluency variable. The average correlation of Fluency with Word Choice is higher at the middle school (.82) and high school (.85) levels than at the elementary level (.75). Fluency also has a stronger average correlation with Conventions at the high school level (.86) than at the lower grades, where the average correlation is .77 in both instances.

The pattern of average correlations among the five Vermont scores for the four middle school sets of papers is quite similar to that shown in Table 5 for the elementary papers. If anything, the average correlations among the first four variables (Purpose, Organization, Details, and Voice) are more uniform at the middle school level. The six average correlations among these four variables range from .82 to .85.

The degree to which a distinction between ratings of overall essay quality and ratings of the more mechanical features of an essay is similar across states was investigated by selecting two scores (an overall quality score and a mechanics or conventions score) for each state that provided ratings on more than one dimension. We selected the Organization score as a proxy for the first four scores related to an overall quality dimension for both Oregon and Vermont. For a second dimension, we selected the Conventions score in Oregon and the Grammar score in Vermont. These pairs of scores were then compared with each other and with the two scores (Content and Mechanics) used by Arizona and at the higher grades with the two scores (Rhetorical Effectiveness and Conventions) used by California.

The overall data collection design at the elementary level yielded two sets of papers that were scored by readers from both states for each of the pairs of states formed by Arizona, Oregon, and Vermont. For example, correlations among the two Arizona scores and the two Oregon scores were computed based on their joint ratings of the papers from Arizona and the papers from New York. Two data sets also were used to compute correlations for the Arizona-Vermont and the Oregon-Vermont pairings.

Tables 6, 7, and 8 report the intercorrelations for the two selected scores from each state for the Arizona-Oregon, Arizona-Vermont, and Oregon-Vermont pairings, respectively. In each table, the correlations based on one set of papers are shown below the diagonal while those for the second set of papers scored by readers from both states are shown above the diagonal. The correlations that are underlined and printed in bold face characters are those that would be expected to be higher than the other listed correlations under the assumption that the Arizona Content, the Oregon Organization, and the Vermont Organization scores are measures of one underlying construct, while the Arizona Mechanics, the Oregon Conventions, and the Vermont Grammar scores are measures of a second common construct. Only one of the six sets of correlations (Table 7, below the diagonal) would meet all the criteria applied to multitrait-multimethod correlation matrices (Campbell & Fiske, 1959) to demonstrate that the trait measures have both convergent and discriminant validity. The criteria would be satisfied partially, however, by the pattern of correlations in two of the other five tables of correlations (Table 6 below the diagonal and Table 7 above the diagonal).

Table 6

Intercorrelations of Elementary Level Arizona Content and Mechanics Scores and Oregon Organization and Conventions Scores (Data Set 1 Correlations Below the Diagonal and Data Set 2 Correlations Above the Diagonal)

Score	Arizona Content	Arizona Mechanics	Oregon Organization	Oregon Conventions
AZ Content	–	.78	.44	.44
AZ Mechanics	.39	–	.49	.48
OR Organization	.82	.56	–	.65
OR Conventions	.43	.53	.70	–

Table 7

Intercorrelations of Elementary Level Arizona Content and Mechanics Scores and Vermont Organization and Grammar Scores (Data Set 1 Correlations Below the Diagonal and Data Set 2 Correlations Above the Diagonal)

Score	Arizona Content	Arizona Mechanics	Vermont Organization	Vermont Grammar
AZ Content	–	.55	.64	.73
AZ Mechanics	.39	–	.24	.56
VT Organization	.74	.60	–	.68
VT Grammar	.39	.71	.69	–

Table 8

Intercorrelations of Elementary Level Oregon Organization and Conventions Scores and Vermont Organization and Grammar Scores (Data Set 1 Correlations Below the Diagonal and Data Set 2 Correlations Above the Diagonal)

Score	Oregon Organization	Oregon Conventions	Vermont Organization	Vermont Grammar
OR Organization	–	.77	.64	.78
OR Conventions	.70	–	.79	.65
VT Organization	.67	.60	–	.69
VT Grammar	.44	.58	.76	–

Table 9

Intercorrelations of Arizona, California, and Selected Oregon and Vermont Scores Based on the Middle School Papers Provided by California

Score	AZ Cont	AZ Mech	CA Rhet	CA Conv	OR Organ	OR Conv	VT Organ	VT Gram
AZ Content	–							
AZ Mechanics	.86	–						
CA Rhet. Effect	.86	.80	–					
CA Conventions	.76	.80	.71	–				
OR Organization	.79	.72	.80	.72	–			
OR Conventions	.80	.81	.78	.77	.72	–		
VT Organization	.77	.74	.73	.60	.76	.64	–	
VT Grammar	.71	.80	.61	.77	.72	.74	.70	–

Note. Cont = Content; Mech = Mechanics; Rhet = Rhetorical Effectiveness; Conv = Conventions; Organ = Organization; Gram = Grammar.

The middle school results for the four sets of ratings of the papers provided by California constitute the most complete example of a test of the convergent and discriminant validity for an overall quality construct and a distinct conventions or mechanics construct. As indicators of the first of these constructs we used Rhetorical Effectiveness for California readers, Content for Arizona readers, and Organization scores for both Oregon and Vermont readers. For the second construct we used Conventions for California and Oregon readers, Mechanics for Arizona readers, and Grammar for Vermont readers.

The intercorrelations of the above eight scores based on the California middle school papers are shown in Table 9. As before, the correlations in bold-faced type and underlined are the ones that should be higher than the other correlations in a given row or column of the matrix to support the claims of convergent and discriminant validity for the two hypothesized constructs. Although there is some weak support for the distinction between two constructs across the four groups of readers, there are a number of exceptions to the pattern.

Cross-state correlations: One score per state. For states with holistic scoring, the comparisons across states are straightforward. For the four states with multiple scores, we computed all possible correlations, but to report all those correlations (1,036 in total) would be quite unwieldy. Furthermore, the results of the analyses of the within-state correlations among the multiple scores and the results of the cross-state analyses in cases of multiple scores that were just summarized suggest that the use of a single score, either a sum of several part scores, or a single selected score, is likely to be just as revealing as a search of the large matrices of intercorrelations.

For the analyses reported below we used the Arizona Content score and the California Rhetorical Effectiveness score because those two scores are conceptually closer to the holistic scores used in other states than are the Mechanics or Conventions scores. For Oregon we used the sum of the Ideas, Organization, Content, and Voice scores, and for Vermont we used the sum of the Purpose, Organization, Details, and Voice scores. Those particular sums were selected, in part, because of the high intercorrelations among those variables and, in part, because the remaining scores for those states (e.g.,

Grammar for Vermont and Word Choice, Sentence Fluency, and Conventions for Oregon) are influenced more by mechanics than the overall quality that is the focus of the holistic ratings in other states. The tendency for the sums selected for Oregon and Vermont to have somewhat higher correlations with holistic scores from other states than the individual scores making up the sums provides some additional support for the decision to create composite scores for those two states.

At the elementary level there were eight sets of papers, one for each of the participating states. Each of these sets was scored by readers from a total of four states (the operational scores obtained from each state when the papers were submitted plus the scores obtained from the three scoring sessions at the workshop). Thus, for each set of papers, six pairs of correlations between refereed scores from readers from one state and refereed scores for readers from another state were computed. This resulted in a total of 48 correlations across the eight sets of elementary level papers.

The 48 cross-state correlations based on a single score per state at the elementary level are displayed in Table 10. Although the table looks like a typical correlation matrix above the diagonal, and like a second one with a few missing entries below the diagonal, it should be kept in mind that the correlations were derived from eight different data sets. Thus, while the table does support observations about the strength of relationship and congruence between the scoring systems of two states defining a cell, comparisons across cells are made somewhat more tentative by the sample-specific nature of each cell.

For most pairs of states there are two correlations based on the joint scoring of two different sets of papers by the readers from that pair of states. For eight of the 28 pairs, however, there is only one correlation because the design only paired those states for one set of papers. For each of the 20 pairs with two correlations, the higher of the two correlations is reported above the diagonal in Table 10 and the lower of the two correlations is reported below the diagonal. For example, the correlations between the Arizona and Colorado refereed scores assigned when readers from both states scored the Colorado elementary paper set is .75, while the corresponding correlation based on the joint scoring of the Arizona elementary paper set is .69. Therefore, the .75 appears above the diagonal and the .69 below the diagonal in Table 10.

Table 10

Elementary Level Cross-State Correlations Using a Single Score per State (Higher Value Above Diagonal and Lower Value Below for Cases Where Correlations for Two Data Sets Were Available)

State	AZ	CO	CT	ME	NY	OR	TX	VT
AZ	–	.75	.71	.76	.73	.82	.72	.74
CO	.69	–	.81	.85	.68	.81	.85	.53
CT	.61	NA	–	.87	.86	.62	.86	.73
ME	NA	.80	.71	–	.75	.69	.87	.80
NY	.52	.57	.69	NA	–	.75	.71	.59
OR	.48	.57	NA	.68	.64	–	.64	.76
TX	NA	.79	.80	.74	.45	NA	–	.78
VT	.67	NA	.73	.64	NA	.74	.77	–

Note. NA = Not available because readers from pairs of states scored only one set of papers in common according to the design shown in Table 2.

From an inspection of Table 10, it is apparent that, with a few exceptions, there is a substantial relationship between the ratings assigned by readers from one state with those assigned by readers of another state. Every state has scores that are correlated .80 or higher with the scores of at least one other state. The median and the mean correlation based on Fisher's Z transformations are both .73 for the 48 correlations in Table 10.

Tables 11 and 12 report the comparable sets of correlations to those in Table 10 for the middle school and high school data sets, respectively. Except for a few low outliers at the middle school level, the cross-state correlations for both the middle school and high school levels tend to be somewhat higher than those at the elementary level, which is not surprising given the general consensus in the field that there is a higher level of agreement in rating papers written by older writers. The median of the 45 correlations for the middle school data in Table 11 is .80 while the mean is somewhat lower (.76) due to the low outliers. The corresponding median and mean for the 42 correlations reported in Table 12 for the high school data sets are .81 and .79, respectively.

Figure 1 provides a more concise summary of the cross-state correlations reported in Tables 10, 11 and 12. (All figures are provided at the end of this document.) The three stem-and-leaf plots show the distributions of the cross-state correlations for the three data sets. In each case the stem shows the tenths digit for a correlation and the leaf shows the hundredths digit. For example, the first row of the elementary level distribution has a stem of .8 and leaves of 5, 5, 6, 6, 7, and 7, which represents six of the elementary level cross-state correlations with values of .85, .85, .86, .86, .87, and .87, respectively.

Although the cross-state correlations displayed in Figure 1 are generally lower than the two-rater reliabilities, it is clear that there is a considerable degree of consensus among the readers from different states in the *relative ordering of written essays* at all three grade levels. The high degree of consistency occurs despite the apparent differences in writing assignments, grade levels (3 to 5, 7-8, or 11-12), and scoring procedures. It is also worth noting that this high level of agreement was achieved without discussions designed to move scoring procedures to a common ground and without any involvement of one state in the choice of a set of essays to score that might be

Table 11

Middle School Level Cross-State Correlations Using a Single Score per State (Higher Value Above Diagonal and Lower Value Below for Cases Where Correlations for Two Data Sets Were Available)

State	AZ	CA	CT	ME	OR	SC	TX	VT
AZ	–	.86	.87	.89	.82	.60	.70	.74
CA	MD	–	.86	.87	.83	.38	.67	.83
CT	.64	NA	–	.88	.65	.75	.74	.88
ME	NA	.83	.86	–	.82	.84	.86	.88
OR	.80	.83	.50	NA	–	.43	.72	.82
SC	MD	NA	.56	.65	NA	–	.87	.87
TX	MD	.37	NA	.34	.32	.50	–	.79
VT	NA	.77	.85	.83	.69	.77	NA	–

Note. MD = Missing data, Arizona middle school operational scores were not received. NA = Not available because readers from pairs of states scored only one set of papers in common according to the design shown in Table 2.

Table 12

High School Level Cross-State Correlations Using a Single Score per State (Higher Value Above Diagonal and Lower Value Below for Cases Where Correlations for Two Data Sets Were Available)

State	AZ	CA	ME	NY	OR	SC	TX
AZ	–	.84	.77	.81	.82	.83	.84
CA	.83	–	.83	.88	.65	.88	.88
ME	.68	.83	–	.87	.87	.83	.84
NY	.73	.86	.71	–	.86	.81	.87
OR	.70	NA	.83	.71	–	.78	.82
SC	.69	.86	.77	NA	.55	–	.73
TX	.59	.74	NA	.78	.80	.66	–

Note. Due to the unbalanced design, three pairs of states read a third set of papers in common. Those additional correlations are: CA-SC, .74; ME-OR, .65; NY-TX, .68. NA = Not available because readers from pairs of states scored only one set of papers in common according to the design shown in Table 2.

Elementary Level

Stem	Leaf	Count	
.8	556677	6	
.8	000112	6	
.7	55566789	8	
.7	1112333444	10	Median = .73
.6	788999	6	Mean* = .73
.6	12444	5	
.5	779	3	75th Percentile = .80
.5	23	2	25th Percentile =.65
.4	58	2	48 Correlations

Middle School Level

Stem	Leaf	Count	
.8	5666677778889	13	
.8	0222333334	10	
.7	5779	4	
.7	0244	4	
.6	5579	4	
.6	04	2	
.5	6	1	Median = .80
.5	00	2	Mean* = .76
.4		0	
.4	3	1	75th Percentile = .86
.3	78	2	25th Percentile = .65
.3	24	2	45 Correlations

High School Level

Stem	Leaf	Count	
.8	666777888	9	
.8	01122333333444	14	Median = .81
.7	7788	4	Mean* = .79
.7	0113344	7	
.6	556889	6	75th Percentile = .84
.6		0	25th Percentile = .71
.5	59	2	42 Correlations

* Means computed using Fisher's Z transformation.

Figure 1. Stem-and-leaf distributions of cross-state correlations.

most compatible with the assignments and scoring procedures used by another state.

Score levels and standards. High cross-state correlations are a necessary, but not a sufficient, condition for claiming that the standards used by different states to evaluate student writing are comparable. High correlations can be achieved despite substantial differences in level of scores assigned or in levels of performance considered excellent, adequate, or failing because the scales all represent a continuum of writing from undeveloped to highly developed. Even if scores assigned by one state were perfectly correlated with those of another, the scores assigned by one state could be uniformly higher than those assigned by the other, or the passing standard could be set at a higher level by one state than by the other.

The establishment of standards is necessarily a matter of judgment (e.g., Jaeger, 1989) and the definition of a common standard will, at a minimum, require considerable opportunity for discussion and negotiation if a broad consensus is to be reached. The most important ingredient is definition of what individual score levels mean: it does not matter what the specific name is, but rather what it represents, in order to compare whether state A has the same standard as state B. Although the cross-state scoring workshop did not attempt to take the step of defining common standards, the results of the workshop do provide information that is relevant for such an effort. Moreover, these data permit some exploration of the extent to which similar standards are implicit in the writing assessments of these several states.

For states that use the same number of scale points (e.g., 1 to 4), some sense of the relative stringency with which the same numerical scale is used may be obtained by comparing distributions of scores for common sets of papers. However, simple comparisons of means or of the percentages of students receiving particular scores are more difficult to make sense of where states use different numerical scales (e.g., 1 to 5, 2 to 12, and 0 to 100).

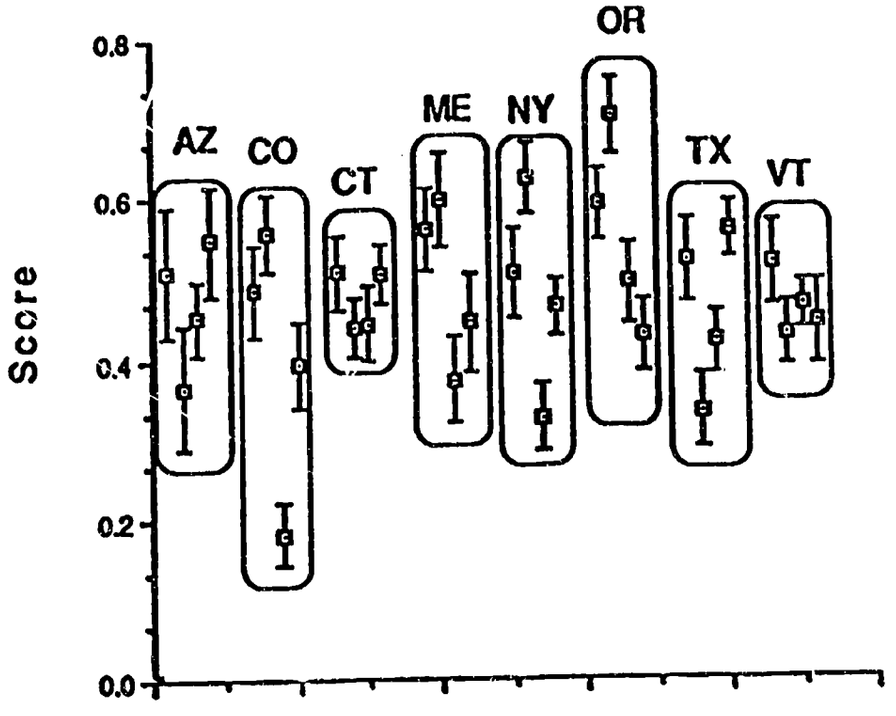
To make the comparison of all states more direct, we converted all the scores to a common scale such that the minimum score was zero and the maximum was one. For example, scores from states using a scale of 1 to 4 were transformed by subtracting 1 and dividing the result by 3. Scores from Maine, which range from 2 to 12, on the other hand, were transformed by

subtracting 2 and dividing the result by 10. A minor modification was used in the case of the New York high school scores, which have a possible range of 0 to 100. Since the lowest score actually assigned to a paper in any of the four sets of papers scored by New York was 10 rather than the allowable minimum of 0, we transformed the New York scores by subtracting 10 and dividing the result by 90.

Means and standard errors of scores in the common 0 to 1 metric are displayed in Figure 2 for the four sets of ratings of all eight elementary paper sets. The four means and their associated standard error bands that were obtained from the scoring of a single set of papers by readers from four states are enclosed in a box. Within each box, the first mean from the left is based on the operational scores obtained from the state providing that set of papers. The next three means, reading from left to right within each box, are based on the scores assigned by the three states that scored that particular set of papers at the workshop. The latter three states are always arranged in alphabetical order from left to right.

For example, the box on the far left of Figure 2 displays the four means and associated standard error bands for the set of papers provided by Arizona. The first mean from the left is based on the operational scores provided by Arizona. The Arizona elementary papers were read by Colorado, Oregon, and Vermont at the workshop. The means based on scores assigned by readers from those three states are represented by the second, third, and fourth points and associated error bands from the left within the first box. In a similar fashion, the means reading from left to right in the box to the far right of Figure 2 are based on the operational scores assigned by Vermont and the scores assigned to the Vermont papers at the workshop by Maine, Oregon, and Texas, respectively.

Three observations based on an inspection of Figure 2 are worthy of note. First, the error bands for papers from a single state (those enclosed within a single box) overlap for all six possible pairs of states doing the scoring in only one of the eight cases (the papers provided by Connecticut). This indicates a fairly substantial diversity in the implicit standards of raters from different states. Second, the mean of the operational scores is the second highest for six states and the highest for the other two states. In all cases, however, the error band for the mean of the operational scores overlaps those associated with the



Notes: From left to right the order of the states scoring the papers provided by each state is as follows:

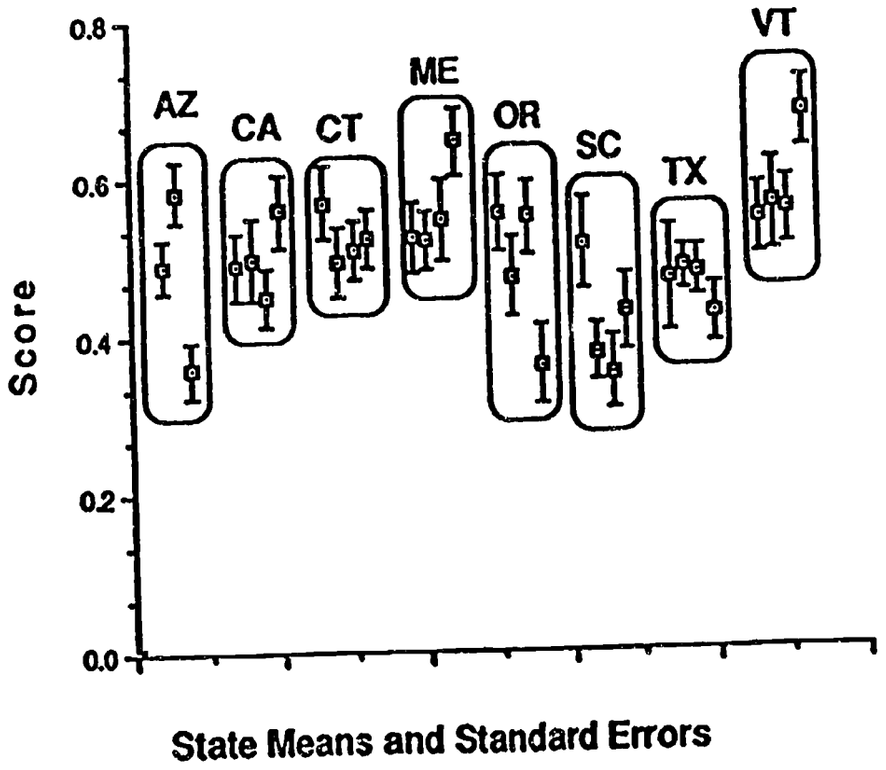
Providing state:	AZ	–	Scoring states:	AZ, CO, OR, VT
	CO	–	Scoring states:	CO, AZ, NY, TX
	CT	–	Scoring states:	CT, AZ, ME, VT
	ME	–	Scoring states:	ME, CO, CT, TX
	NY	–	Scoring states:	NY, AZ, CT, OR
	OR	–	Scoring states:	OR, CO, ME, NY
	TX	–	Scoring states:	TX, CT, NY, VT
	VT	–	Scoring states:	VT, ME, OR, TX

Figure 2. Elementary level mean scores and standard errors by sending state and state reading papers.

means obtained from readers from at least one, and usually more than one, of the other states during the workshop. Finally, there is a fairly clear order of states in terms of the mean scores obtained from readings at the workshop. Ignoring the means for the operational scores, it can be seen that, for Connecticut and New York readers, means are consistently lower than those of readers from the other states. Vermont, Oregon, and Colorado, on the other hand, have means consistently higher than those of other states, while Arizona, Maine, and Texas fall in between.

Means and standard errors for middle schools on the common 0 to 1 metric are displayed in Figure 3 in a manner parallel to the Figure 2 display of the elementary level results. Perhaps the most obvious contrast between the middle school and elementary level results is that there is generally greater overlap of the middle school standard error bands, not only within most of the boxes for the papers from a single providing state, but across paper sets as well. The greater overlap occurs despite a tendency for the error bands to be shorter at the middle school level than at the elementary level. Since the number of papers provided by states was the same at the two lower grade levels, the shorter error bands at the middle school level reflect generally smaller standard deviations (fewer minimum or maximum scores) at the middle school level than at the elementary level.

It should be noted that the operational scores for the Arizona middle school papers were not provided. Hence, the first box on the extreme left of Figure 3 contains only the means for the three states that read the Arizona papers at the workshop. The tendency for means of the elementary level operational scores to be second highest or highest was not observed for the seven middle school states with operational scores. Instead, the operational mean is highest in three cases (Connecticut, Oregon, and South Carolina), second lowest in three cases (California, Maine, and Texas), and lowest in the remaining case (Vermont). In keeping with the greater overlap of error bands for the middle school results, there is not as clear an indication that some states are scoring consistently higher or lower than others for the middle school papers as was noted for the elementary papers. Focusing only on scores assigned at the workshop, the means of scores assigned by South Carolina readers are higher than those assigned by other states. There is less consistency in the relative ordering of other states, however. Means of Texas



State Means and Standard Errors

Notes: From left to right the order of the states scoring the papers provided by each state is as follows:

Providing state:			
AZ	–	Scoring states:	CA, SC, TX
CA	–	Scoring states:	CA, AZ, OR, VT
CT	–	Scoring states:	CT, AZ, ME, SC
ME	–	Scoring states:	ME, CA, CT, VT
OR	–	Scoring states:	OR, AZ, CT, TX
SC	–	Scoring states:	SC, ME, TX, VT
TX	–	Scoring states:	TX, CA, ME, OR
VT	–	Scoring states:	VT, CT, OR, SC

Figure 3. Middle school mean scores and standard errors by sending state and state reading papers.

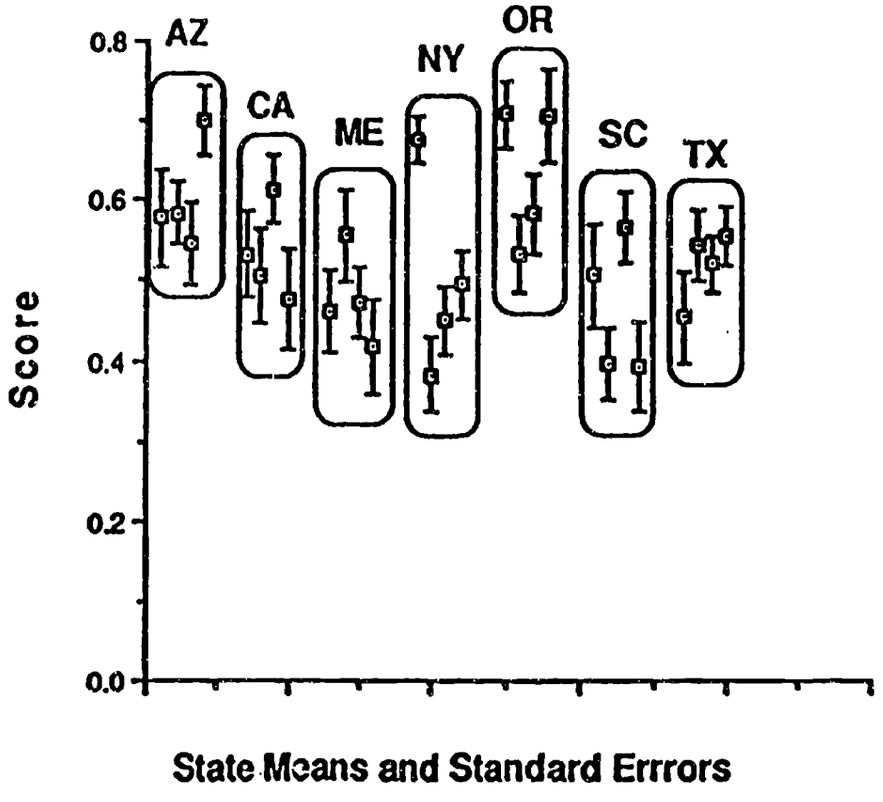
readers, for example, are noticeably low compared to California or South Carolina readers for the Arizona papers and to Arizona or Connecticut readers for the Oregon papers, but high compared to Arizona or Maine readers for the California papers.

High school level results on the common 0 to 1 metric are displayed in Figure 4. The spread of means for a single paper set, as well as across paper sets, appears more similar to the elementary level results than to the middle school results. The larger error bands for high school papers reflect, in part, smaller sample sizes. Recall that states were asked to provide only 36 papers at the high school level as compared to 45 at the two lower grade levels.

New York readers tend to have the highest means on the common metric in Figure 4. The score scale used by New York at the high school level, 0 to 100, is quite different in character from the scales used by other states, however. Thus, the apparently high means for New York in Figure 4 may be an artifact of the type of scale that leads to greater use of the top half of the score range than of the bottom half. The notion that a 70 is a passing score makes a 50 a low score, for example.

As was true at the middle school level, there is not a clear tendency for the means of operational scores to be higher or lower than those obtained from readers from other states. Nor is there a clear pattern (with the possibly artifactual exception of New York) indicating that high school level readers from one state are consistently more stringent or more lenient than their counterparts from other states.

Comparisons of means are a useful first step, but leave unanswered other questions about distributions of scores or how comparable a pass-fail dichotomy or some other dichotomy based on a common standard would be. Simple contingency tables can be used to shed additional light on these issues. Contingency tables provide not only a means of displaying the level of agreement reflected in the correlations, but also a way of comparing the relative stringency of the ratings across states. Such comparisons are most straightforward where the states use the same number of scale points. They are likely to be most informative when the states also use cutting points to define similar actions (e.g., the need for remediation, or a minimum requirement for graduation). In other cases, however, the contingency table



Notes: From left to right the order of the states scoring the papers provided by each state is as follows:

Providing state:				
	AZ	–	Scoring states:	AZ, OR, SC, TX
	CA	–	Scoring states:	CA, AZ, NY, TX
	ME	–	Scoring states:	ME, AZ, CA, SC
	NY	–	Scoring states:	NY, AZ, ME, OR
	OR	–	Scoring states:	OR, CA, ME, SC
	SC	–	Scoring states:	SC, CA, NY, TX
	TX	–	Scoring states:	TX, ME, NY, OR

Figure 4. High school mean scores and standard errors by sending state and state reading papers.

results may still provide the basis for discussing why one state assigns relatively few papers the highest possible score while another state assigns their highest score to many more of the same set of papers. This may be a result of the different purposes of the state writing scales. Some may be based on a minimum-competency approach while others are based on a broader range of development with less of a ceiling effect. An example of this distinction is the fact that the New York papers and criteria do not represent those used for the Regents examination that is required for the more prestigious Regents Diploma.

One difficulty of contingency tables for present purposes is the sheer number of possibilities. It would require 132 contingency tables, one for each of the correlations reported in Tables 10, 11, and 12, to illustrate all of the pairwise comparisons at the three grade levels. To reduce this number to something more manageable, we selected three tables per grade level to report here. The correlations summarized in Figure 1 were used to select a pair of states that had a correlation equal to the 75th percentile, the median, and the 25th percentile in the distributions of correlations at each grade level (see Figure 1). Where more than one pair of states had a correlation equal to one of these points in the distribution, we selected pairs so that all ten states would appear in at least one of the contingency tables and no state would appear in more than three of them.

Table 13 presents the contingency table corresponding to a correlation of .80 which is at the 75th percentile of the distribution of cross-state correlations at the elementary level. Scores assigned by readers from Maine are displayed as rows and those assigned by readers from Colorado as columns. The Maine scores are the sum of two ratings on a 1 to 6 scale. Hence, they range from 2 to 12 while the Colorado scores are on a 1 to 4 scale. In other data sets, the Maine scores include odd numbers such as 11 where one rater assigned a 5 and the other, a 6. For this set of papers, however, the refereed scores were simply doubled, so there are no odd-numbered scores.

A contingency table such as that shown in Table 13 could be used as the basis of a discussion across states of the comparability of different standards. For example, if a score of 4 was considered excellent by Colorado, what Maine score would most nearly correspond? Sixteen papers received a score of 4 from the Colorado raters, while 2, 8, and 18 of the papers received score of 12, 10 or

Table 13

Cross-Tabulation of Maine and Colorado Elementary Scores (Correlation is .80 and at the 75th Percentile)

		Colorado Elementary Scores				
		1	2	3	4	Row Total
Maine Elementary Scores	12				2	2 4.8
	10			1	5	6 14.3
	8			4	6	10 23.8
	6		3	8	3	14 33.3
	4	2	5			7 16.7
	2	2	1			3 7.1
	Column Total	4 9.5	9 21.4	13 31.0	16 38.1	42 100.00

higher, and 8 or higher, respectively, from the Maine readers. In other words, a Colorado 4 would appear to be less demanding than a Maine 10 or 12, and roughly comparable to a Maine score of 8 or higher for these data. If Maine were to require a minimum score of 6 to pass, 10 of the papers reflected in Table 13 would fail to meet that standard. On the other hand, if Colorado required a 3 as the minimum passing grade there would be 13 failures.

Tables 14 and 15 present the other two example contingency tables for the elementary level data. Table 14 reflects data with a correlation approximately at the median of the distribution of cross-state correlations while Table 15 reflects a correlation at the 25th percentile. It can be seen in Table 14 that a Connecticut score of 3 or 4 most closely corresponds to a paper that was rated "Frequently" by both Vermont readers, "Frequently" by one and "Extensively" by the other, or "Extensively" by both. However, a total of 14 of the 45 papers reflected in Table 14 received ratings of Frequently or Extensively from the Vermont readers, whereas 11 of the papers received scores of 3 or 4 from the Connecticut readers.

Table 15 provides a clear illustration of apparently different degrees of leniency in the use of what is nominally the same 1 to 4 point scale. The Colorado readers assigned scores of 4 to 16 of the 40 papers, whereas New York readers assigned scores of 4 to only 3 of the same papers. Indeed, there is one fewer paper with a score of either 3 or 4 according to the New York readers than there are papers with scores of 4 according to the Colorado readers. If a score of 3 or higher was required to meet a mastery standard in both states, 25 of the 40 papers would fail to meet that standard according to the New York readers, whereas only 11 would fail to meet the standard according to the Colorado readers.

Tables 16, 17, and 18 present the cross-tabulations of three middle school pairs of scores from different states that have correlations at the 75th percentile, the median, and the 25th percentile, respectively. As can be seen in Table 16, not only is the relationship between the California Rhetorical Effectiveness scores and the Arizona Content scores quite strong, but there is a relatively simple mapping of one set of scores into the other that would yield comparable standards at the upper end of the score scales. A standard of a California score of 5 or higher would correspond quite closely to an Arizona

Table 14

Cross-Tabulation of Vermont Elementary Details Scores and Connecticut Elementary Scores (Correlation is .74, Approximately at the Median)

Vermont Elementary Details Scores	Connecticut Elementary Scores 1	2	3	4	Row Total
Extensively			1		1 2.2
Extensively/ Frequently			2	2	4 8.9
Frequently	1	3	4	1	9 20.0
Frequently/ Sometimes	2	6	1		9 20.0
Sometimes	6	8			14 31.1
Sometimes/ Rarely	5	1			6 13.3
Rarely	2				2 4.4
Column Total	16 35.6	18 40.0	8 17.8	3 6.7	45 100.0

Table 15

Cross-Tabulation of Colorado and New York Elementary Scores (Correlation is .68 and at the 25th Percentile)

Colorado Elementary Scores	New York Elementary Scores 1	2	3	4	Row Total
4		3	12	1	16 40.0
3	2	9	1	1	13 32.5
2	2	7			9 22.5
1	2				2 5.0
Column Total	6 15.0	19 47.5	13 32.5	2 5.0	40 100.00

Table 16

Cross-Tabulation of California Rhetorical Effectiveness Scores and Arizona Content Scores for the California Middle School Papers (Correlation is .86 and is at the 75th Percentile)

California Rhetorical Effectiveness Scores	Arizona Content Scores 1	2	3	4	Row Total
6			1	4	5 11.1
5			1	4	5 11.1
4		4	5	1	10 22.2
3		7	3		10 22.2
2	4	7			11 24.4
1	4				4 8.9
Column Total	8 17.8	18 40.0	10 22.2	9 20.0	45 100.00

Table 17

Cross-Tabulation of Oregon Ideas Scores and Arizona Content Scores for the Oregon Middle School Papers (Correlation is .78 and is Approximately at the Median)

Oregon Ideas Middle School Scores	Arizona Content Scores 0	1	2	3	4	Row Total
10				3	2	5 13.5
9				3	1	4 10.8
8			2	2		4 10.8
7			2	4		6 16.2
6			2	3		5 13.5
5	1	1	3	1		6 16.2
4		2	1			3 8.1
3		3				3 8.1
2		1				1 2.7
Column Total	1 2.7	7 18.9	10 27.0	16 43.2	3 8.1	37 100.00

Table 18

Cross-Tabulation of Oregon Organization and Connecticut Scores for Vermont Middle School Papers (Correlation is .65 and at the 25th Percentile)

Oregon Organization Middle School Scores	Connecticut Middle School Scores 1	2	3	4	Row Total
10				4	4 8.9
9			1	4	5 11.1
8			4	3	7 15.6
7		5	2		7 15.6
6	1	2	4	1	8 17.8
5			2	1	3 6.7
4	4	3			7 15.6
3	2	1			3 6.7
2				1	1 2.2
Column Total	7 15.6	11 24.4	13 28.9	14 31.1	45 100.00

score of 4, and California scores of 4 or higher would correspond well to Arizona scores of 3 or higher.

The larger number of scale points for the Oregon scores and the weaker relationships with the other scores in Tables 17 and 18 make the mapping of a standard for one state into a comparable one for the other state weaker than would be possible for the Table 16 results. The Oregon-Arizona comparison (Table 17) shows that an Oregon score of 7 or higher would correspond to an Arizona score of 3 or higher in that those standards would each qualify 19 of the papers. Furthermore, 15 of the 19 papers that meet that standard of 7 for Oregon also meet the standard of 3 for Arizona. Other possible standards would correspond less well, however. Possible choices of standards for the Oregon-Connecticut comparison shown in Table 18 would correspond less well.

The final set of cross-tabulations are displayed in Tables 19, 20, and 21 for the three selected pairs of high school scores. There is a fairly close correspondence between Arizona and Texas scores of 3 or higher (Table 19). Fifteen of the papers reflected in Table 20 have New York scores of 70 or higher and 18 have South Carolina scores of 3 or higher. If 3 and 70 were selected as common standards for South Carolina and New York, respectively, 13 of the papers would meet the standards according to readers from both states, 2 would meet the standard for New York but not South Carolina, and 5 would meet the standard for South Carolina but not New York. Twelve of the 36 papers summarized in Table 21 received scores of 70 or higher from New York raters and 12 of the papers received scores of 8 or higher from the Oregon raters. If 70 and 8 were used as standards for New York and Oregon respectively for the high school papers summarized in Table 21, then 8 papers would meet the standards of both.

Discussion and Conclusions

The results of the cross-state scoring workshop indicate that a substantial consensus already exists regarding what defines the relative quality of student writing at the elementary, middle, and high school levels. The high correlations between the scores assigned by readers from different states were achieved without modification of the scoring procedures used by the ten states that participated in the workshop. Nor was there any attempt to select writing

Table 19

Cross-Tabulation of Arizona Content and Texas Scores for California High School Papers (Correlation is .84 and at the 75th Percentile)

		Texas High School Scores				
		1	2	3	4	Row Total
Arizona Content High School Scores	4			3	6	9 25.7
	3		2	3	1	6 17.1
	2	3	7	4		14 40.0
	1	6				6 17.1
	Column Total	9 25.7	9 25.7	10 28.6	7 20.0	35 100.00

Table 20

Cross-Tabulation of New York and South Carolina High School Scores
(Correlation is .81 and at Approximately the Median)

New York High School Scores	South Carolina Scores 1	2	3	4	Row Total
100.00				1	1 2.8
90-99		1		4	5 13.9
80-89			2	2	4 11.1
70-79		1	3	1	5 13.9
60-69		5	4		9 25.0
50-59	1			1	2 5.6
40-49	3				3 8.3
30-39	1	2			3 8.3
20-29	3				3 8.3
10-19	1				1 2.8
Column Total	9 25.0	9 25.0	9 25.0	9 25.0	36 100.00

Table 21

Cross-Tabulation of New York and Oregon Content Scores for Texas High School Papers (Correlation is .68 and at Approximately the 25th Percentile)

		Oregon High School Content Scores								
		2	4	5	6	7	8	9	10	Row Total
New York High School Scores	90-99							1		1 2.8
	80-89						2	1	1	4 11.1
	70-79				2	2		3		7 19.4
	60-69			1	1		1		1	4 11.1
	50-59			1	3	3	1	1		9 25.0
	40-49		1	2	2	1				6 16.7
	30-39		1	1	1	1				4 11.1
	10-19	1								1 2.8
	Column Total	1 2.8	2 5.6	5 13.9	9 25.0	7 19.4	4 11.1	6 16.7	2 5.6	36 100.00

prompts or assignments that were most compatible with those used by other states. In view of the substantial differences across states in the nature of the writing assignments, the uses that are made of the scores, the scoring procedures, and the age of the state writing assessment programs, the high level of agreement that was achieved provides good documentation of the fact that these states share a common view regarding the relative ordering of student writing from low to high. Even though the scales may not all represent the full range of writing competency, the criteria of poor to good writing is ordered in quite similar ways across the 10 participating states.

Since our focus was on the agreement of readers across states, only one writing sample was obtained per student. It should be noted, however, that previous research has shown that consistency of scores across raters is generally a good deal better than consistency of scores that a given group of students receives when multiple writing samples are scored for each student (Breland, Camp, Jones, Morris, & Rock, 1987; Coffman, 1966; Dunbar, Koretz, & Hoover, in press; Hieronymus & Hoover, 1986). This suggests that it will be important in the development of a procedure for comparing state results to a national standard to consider multiple samples of writing from each student.

The high level of agreement on the relative ordering of papers by readers from different states that was achieved in this study is a prerequisite for any system that would compare results within a state to a common national standard. But the agreement on relative order of papers is not the same as agreeing on absolute judgments in comparison to a set standard of performance. There were some rather large differences in the level of scores that were assigned to papers even in instances where the relative ordering of papers by different states was almost as similar as the within-state rating reliabilities would allow. Even states that use the same number of score points, most commonly 1 to 4, sometimes differed substantially in the implicit standard with which the readers assigned the scores. In extreme cases, it appears to be about as difficult for a paper to earn a score of 3 according to the standards of one state as it is to earn a 4 according to the standards of another state.

Such differences in apparent stringency of scoring do not preclude the possibility of agreeing on a common standard. In the above extreme case, for example, it is conceivable that a score of 4 would be required as the standard of

excellence in the state with the more lenient scoring, whereas a 3 or higher would be required in the state with the more stringent scoring. It seems clear, however, that considerable cross-state discussion would be required to arrive at common performance standards. Each state delegation will need to review their scoring rubrics to decide if their scale currently reflects a high standard at the top end. The criteria for determining the high standard need to be specifically stated in order to ascertain if there is agreement on the standard itself, regardless of the specific number assigned by any given state.

The vision of the New Standards Project, which requires specifically defined criteria upon which a student will be judged, is directly related to the need for the state writing personnel to review their scoring scales. Although a single score may be easier to assign, it may be that multiple scores better communicate the criteria on which a student will be measured in meeting a standard.

References

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M., & Rock D.A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examination Board.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Coffman, W.E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3*, 151-156.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (in press). Quality control in the development and use of performance assessments. *Applied Measurement in Education.*

Hieronymus, A.N., & Hoover, H.D. (1986). *Iowa Tests of Basic Skills: Writing supplement teacher's guide*. Chicago: Riverside Publishing Co.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 485-514). New York: Macmillan.

LRDC & NCEE. (no date). *The New Standards Project: An overview.* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

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

Resnick, L.B. (1991, April). *Examinations and learning*. Paper presented in a symposium entitled "Setting a New Standard: Toward an Examination System for the United States" presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Tucker, M. (1991, April). *National examinations and the education reform agenda*. Paper presented in a symposium entitled "Setting a New Standard: Toward an Examination System for the United States" presented at the annual meeting of the American Educational Research Association, Chicago, IL.