C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Cross-scorer and Cross-method Comparability and Distribution of Judgments of Student Math, Reading, and Writing Performance

Results From the New Standards Project Big Sky Scoring Conference

CSE Technical Report 368

Lauren Resnick and Daniel Resnick
CRESST/LRDC, University of Pittsburgh

Lizanne DeStefano
University of Illinois at Urbana-Champaign

November 1993

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

The work reported herein was supported in part 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.

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.

CROSS-SCORER AND CROSS-METHOD COMPARABILITY AND DISTRIBUTION OF JUDGMENTS OF STUDENT MATH, READING, AND WRITING PERFORMANCE

RESULTS FROM THE NEW STANDARDS PROJECT BIG SKY SCORING CONFERENCE

Lauren Resnick and Daniel Resnick
CRESST/Learning Research and Development Center,
University of Pittsburgh

Lizanne DeStefano
University of Illinois at Urbana-Champaign

BACKGROUND

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 and is funded in part by the National Center for Research on Evaluation, Standards, and Student Testing.[1] The project is an effort to create a state- and district-based assessment and professional development system that can serve as a catalyst for major educational reform. As part of a professional development strategy tied to assessment, 114 teachers, curriculum supervisors, and assessment directors, representing 23 states and districts, met in Big Sky, Montana on June 27 through July 1, 1992 to refine rubrics and procedures and to score student responses. The responses were collected in the spring 1992 field test of mathematics and English language arts performance tasks administered to close to 10,000 fourth-grade pupils in the partner states and districts.

[1] The National Center for Research on Evaluation, Standards, and Student Testing is 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. 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.

The Spring 1992 Field Test

The spring 1992 field test was part of this development effort rather than a rigorous pilot of the fully developed on-demand assessment component of New Standards. More than 400 teachers and their students were involved in the field test pilot, but teacher participation was voluntary and the choice of which classes to test was left to the teacher. No attempt was made to systematically select groups of students on the basis of ability, demographics, or geography. As a result, while the data produced by the field test tell us much about the characteristics of individual tasks and the scoring processes in general, they are not representative of any well-defined group of students, and should not be used to make judgments about student abilities or the nature of classrooms.

The Nature of Tasks Scored

Tasks were scored in reading, writing, and mathematics. In reading, three tasks, Folk Tales (#17), Camels (#4), and Wolves (#3), were scored using two methods. In the first method, *holistic scoring*, a single score was assigned to represent the quality of student response across all items associated with a task. The second method, *anaholistic scoring*, required scorers to assign individual scores to each item (analytic) and then to assign an "Across Item" score (holistic). A third type of anaholistic score was produced by summing the individual item scores. Two tasks, Camels (#4) and Folk Tales (#17), were scored using both methods so that cross-method comparisons could be made. Both holistic and anaholistic scoring methods in reading used a 5-point rating scale (0–4).

Three tasks were also scored in writing: Memories (#18), Camels (#4), and Wolves (#3). Scoring in writing was holistic, using a 7-point (0–6) rating scale format. Two tasks, Camels (#4) and Wolves (#3), were scored for both reading and writing so that cross-area comparisons could be made. Scorers were trained to score both reading and writing tasks during the Big Sky Scoring Conference.

In mathematics, seven tasks were scored: Amusement Park (#8); Name that Graph (#9); Hot Dog Rolls (#10); Pizza Party (#12); Building with Tiles (#15); Too High, Too Low (#21); and The Aquarium Problem (#73). Math

scoring used holistic assignment of scores according to a 7-point (0–6) rating scale.

PROCEDURE

At Big Sky, scorers were acquainted with rubrics and exemplars, trained in the scoring process, and asked to score sets of papers during a four-day period. Given that it included both reading and writing, training for English language arts scoring was severely limited by time. A half-day only was devoted to the scoring of a single task in reading or writing. Each half-day session began with 60–90 minutes spent reviewing the task, the rubric, and exemplars of student work at each score level. This was followed by work on a common training set, followed by discussion of disputed scorings. Then independent scoring commenced. In the first day and a half, three tasks were scored and double scored in reading. In the next day and a half, the same set of scorers scored and double scored three tasks in writing.

The training in mathematics was more rigorous and provided more continuous feedback to scorers through a calibration process. The first half-day was spent discussing the agenda and preparing to score the first task by performing the task, discussing and debating the rubric levels, drafting a rubric, comparing the draft rubric to the New Standards Project (NSP) rubric, and reflecting on the results. The remainder of the day was allocated for calibration and scoring of the first task. The calibration process started with an examination of the benchmark papers. The exemplar papers then were scored a few at a time and the scores discussed until each member of a table was in agreement as to which scores the responses should receive. After training, each person scored independently. Approximately 10% to 15% of the papers were double scored by the table leader or another member of the group. For each additional task, the cycle of performing the task, discussing the rubric, scoring the exemplar papers, calibrating the table to produce the same scores on each paper, scoring independently and double scoring was repeated. Over a three-day period, seven tasks were scored in mathematics.

A total of 2,178 student booklets (across three tasks) were scored for reading, 2,206 student booklets (across three tasks) were scored for writing, and 5,084 student booklets (across seven tasks) were scored for math.

Recording of Data

Throughout the scoring conference, as booklets were scored, scoring sheets were collected and data were entered into *Excel* spreadsheets. Demographic information and student performance files were converted to *SYSTAT* for analysis.

Double Scoring

For all tasks in reading, writing, and mathematics, 10% to 20% of all student responses were randomly selected for scoring by two different scorers. Data from double scoring were used to assess cross-scorer comparability of scores.

About 50% of student responses to two reading tasks, Camels (#4) and Folk Tales (#17), were randomly selected and scored using both holistic and anaholistic methods. Data from double scoring were used to assess the cross-method comparability of scores.

FINDINGS

Comparability Across Scorers

Interscorer comparability was represented in three ways: (a) correlation coefficients between first round and second round scores for double-scored responses; (b) exact percent agreement between first and second round scores (i.e., the percent of responses where the same score was assigned in both rounds); and (c) adjacent percent agreement between first and second round scores (i.e., the percent of responses in which first and second round scores differed by one point). Because it may be lowered by the restricted range of score options, the correlation coefficient (like exact percent agreement) can be considered to be the most conservative measure of comparability. In contrast, adjacent percent agreement allows considerable cross-scorer fluctuation in scoring and, especially in the case of limited scoring ranges, may grossly inflate comparability. For these reasons, none of these indicators taken alone is a sufficient indicator of cross-scorer comparability. Taken together, the three indicators provide a fuller picture of the stability of scores across different scorers and the fidelity of application of rubrics.

Reading

Table 1 reports correlations, percentage exact agreement and percentage adjacent agreement for the three reading tasks. Results are reported for both anaholistic and holistic scoring methods for Folk Tales (#17) and Camels (#4). Only holistic scoring was done for Wolves (#3).

Correlations for holistic scoring. Reliability coefficients for the holistically scored tasks were moderate, ranging from .47 for Camels (#4) to .67 for Folk Tales (#17). These coefficients are considerably lower than levels obtained with traditional, standardized assessments and seriously limit the use of

Table 1

Indicators of Interscorer Comparability for Three Reading Tasks: Folk Tales, Camels, and Wolves

Task		Correlation	% Agreement (exact)	% Agreement (adjacent; ±1pt)
Folk Tales (#17)				
Holistic (*N* = 121)		.67	62%	99%
Anaholistic (*N* =126)	Item 1	.75	72%	88%
	Item 2	.67	66%	88%
	Item 3	.87	84%	96%
	Item 4	.78	56%	88%
	Item 5–6	.81	72%	88%
	Item 7	.44	58%	85%
	Item 8	.45	35%	88%
	Across	.81	88%	98%
	Sum	.90		
Camels (#4)				
Holistic (*N* = 82)		.47	54%	89%
Anaholistic (*N* =146)	Item 1	.65	63%	89%
	Item 2	.64	63%	89%
	Item 3	.69	72%	92%
	Item 4	.59	59%	92%
	Item 5	.72	65%	91%
	Across	.71	62%	87%
	Sum	.79		
Wolves (#3)				
Holistic (*N* = 124)		.59	55%	99%

holistic reading scores from the 1992 spring pilot. They are not so low, however, as to suggest that different scorers cannot be trained to reach a high level of agreement on performance tasks such as those in the pilot. In fact, given the limited amount of training and pilot nature of the tasks and the rubrics, it is reasonable to expect that greater reliability for holistic scoring is attainable.

Correlationsforanaholisticscoring. Coefficients for individual items scored using the anaholistic method covered a broad range, from a low of .44 (Item 7–Camels) to a high of .87 (Item 3–Folk Tales). Four out of the 12 item scores had reliabilities equaling or exceeding .75, a commonly used lower limit for acceptable levels of reliability. It may be useful to examine the characteristics of the rubrics, exemplars, and scorer training for those items, as well as the items themselves, to ascertain characteristics that may have positively influenced scorer agreement for those items. Likewise, by examining items with low coefficients, we may identify aspects of scoring that mitigate against agreement.

Anaholistic "Across" scores were holistic scores (range = 0 to 4) assigned to the student's response after the item scores were assigned. These scores were more reliable (Folk Tales = .81 and Camels = .71) across scorers than the holistic scores assigned without individual item ratings (.67 and .47, respectively). This finding supports the use of anaholistic scoring over holistic scoring by suggesting that scorers' agreement is aided by an item-by-item assessment prior to assigning a holistic score.

Anaholistic "Sum" scores, produced by summing the individual item scores, proved to be the most reliable across scorers, with both coefficients (Folk Tales = .90 and Camels = .79) in the acceptable range. This latter finding is not surprising, given our earlier comment that reliability coefficients are directly influenced by the spread of scores in the group tested. Because larger reliability coefficients result when individuals tend to stay in the same relative position in a group from one testing to another, it naturally follows that anything that reduces the possibility of changing positions in the group contributes to higher reliability coefficients. Greater differences between the scores of individuals reduce the possibility of changing positions (Gronlund & Linn, 1990). In other words, with other things being equal, the larger the spread of scores, the higher the estimate of reliability will be. In the case of

holistic scores, anaholistic item scores, and "Across" scores, the scores are confined to a range of 5 score points (0–4). The "Sum" score, created by adding all the item scores, can range from 0 to 4 × (the number of items on a task), in the case of Camels (#4) from 0 to 20.

Although they may differ in scale, the "Sum" and the "Across" scores are similar in that they represent holistic scores that are based on analysis of smaller parts of the response. They are highly related to each other; their intercorrelations ranged from .68 to .86 across the three reading tasks. This indicates that either may be used as a global measure of the quality of student performance, but the higher interscorer reliability and increased range of the "Sum" score may make it more attractive for use as a composite score.

Exact percent agreement. Another way to represent consistency across scorers that may be clearer to nontechnical audiences is the percentage of times that two different scorers assign the exact same score to a student's response. Looking at Table 1, exact percent agreement figures present a picture similar to that portrayed by the correlations. When assigning holistic scores, scorers agreed more than half the time. Exact agreement when assigning individual item scores in the anaholistic method varied widely. "Across" scores had higher percentages of exact agreement than holistic scores for the same tasks. Because of the wide range of "Sum" scores, exact agreement percentages were not reported.

Adjacent percent agreement. It is argued by some that exact agreement is too stringent a condition to impose upon a reliability estimate. Percentage adjacent agreement classifies ratings that are within one point of each other as "agreement." The third column of Table 1 presents adjacent agreement percentages for the three reading tasks. It is not surprising that adjacent agreement percentages present a much rosier picture of the extent to which raters agree than the other, more conservative measures. It is important to note that adjacent percent agreement was reported here to provide information about the scoring process and not as a measure of reliability. On a 5-point scale, even random assignment of scores will produce a large percentage of cases in which scorers are within one point of each other. Therefore, adjacent percent agreement is not an acceptable indicator of reliability for the 1992 spring pilot tasks.

Analysis of bivariate distributions comparing scores assigned by first and second scorers (Tables 2, 3, and 4) revealed that the difference between first and second scoring was most often one point and that lack of agreement occurred most frequently on papers with midrange scores (3s on a 0 to 4 scale). Scorers seemed to most consistently agree on scores for papers in the lower score range (1s and 2s on a 0 to 4 scale). There were too few high scoring papers in the double-scored sample to comment upon agreement at the upper end of the score range.

Table 2

Percent Agreement Comparison of Holistic Reading Scores From First and Second Scorings for Folk Tales (#17) (*N*=121)

First score	Second score 1	2	3	4	Total %	*N*
1	73.17	24.39	2.44	0.00	100.00	41
2	12.00	60.00	24.00	4.00	100.00	50
3	0.00	37.93	51.72	10.34	100.00	29
4	0.00	0.00	100.00	0.00	100.00	1
N	35	51	29	5		

Table 3

Percent Agreement Comparison of Holistic Reading Scores From First and Second Scorings for Camels (#4) (*N*=82)

First score	Second score 1	2	3	4	Total %	*N*
1	65.52	20.69	13.79	0.00	100.00	29
2	21.05	52.63	21.05	5.26	100.00	38
3	25.00	25.00	33.33	16.67	100.00	12
4	0.00	0.00	66.67	33.33	100.00	3
N	30	29	18	5		

Table 4

Percent Agreement Comparison of Holistic Reading Scores From First and Second Scorings for Wolves (#3) (*N*=123)

First score	Second score 1	2	3	4	Total %	*N*
1	61.29	32.26	6.45	0.00	100.00	31
2	28.33	56.67	15.00	0.00	100.00	60
3	3.57	46.43	46.43	3.57	100.00	28
4	0.00	0.00	60.00	40.00	100.00	5
N	37	57	27	3		

These findings stress the need to create rubrics that clearly differentiate across all score values and to provide exemplars that illustrate distinctions across score values, especially at the middle of the scale. If rubrics and exemplars are adequate and their reliability is still low at certain score values, it may be that the score scale should be reduced, collapsing values in the middle of the scale where agreement is difficult to achieve.

Writing

Correlations, percentage exact agreement, and percentage adjacent agreement for the three writing tasks, Memories (#18), Camels (#4), and Wolves (#3), are reported in Table 5.

Table 5

Indicators of Interscorer Comparability for Three Writing Tasks: Memories, Camels, and Wolves

Task	*N*	Correlation	% Agreement (exact)	% Agreement (adjacent; ±1pt)
Memories (#18)	439	.53	49%	86%
Camels (#4)	237	.56	42%	88%
Wolves (#3)	705	.54	41%	86%

Correlations. Holistic scoring of standardized performance assessments in writing has become commonplace in large-scale state and national assessment efforts. With its increased popularity, improved procedures for scoring and training scorers have resulted in greater technical adequacy of these assessments. For example, it is not unusual to see reliability coefficients of .70 or greater associated with standardized performance assessments in writing (Vermont Department of Education, 1991). Across the 1992 New Standards spring pilot tasks, correlations between first and second round scores for writing were consistent (Range: Memories = .53 to Camels = .56), but below the range generally considered acceptable. As in reading, these low reliabilities limit the use of writing data from the 1992 spring pilot. They are not so low, however, to suggest that interscorer reliability is not attainable, especially given that other large-scale assessments have had considerable success. Revision of training materials and processes seems warranted—especially reconsideration of training people to score both reading and writing in a single conference. On the conference evaluation survey, scorers for the literacy tasks reported feeling fatigued and overwhelmed at having to score both reading and writing tasks.

Exact percent agreement. When assigning holistic scores on a 7-point scale (0–6), scorers agreed less than half the time. Exact percent agreement for the writing tasks averaged 47% (Range: Wolves = 41% to Memories = 49%). Exact agreement was somewhat lower for writing than for the reading tasks, possibly an artifact of the 7-point scale in writing allowing for more disagreement than the 5-point scale used in reading.

Adjacent percent agreement. As in reading, most scores varied by only one point between the first and second round (adjacent percent agreement ranged from 86% to 88%). As illustrated in Tables 6, 7, and 8, once again, scoring shifts were most likely to have occurred in midrange papers (3s, 4s and 5s on a 0 to 6 scale). Scorers seemed to most consistently agree on scores for papers in the lower score range (1s and 2s on a 0 to 6 scale). Once again, there were too few papers scored "6" in the double-scored sample to assess agreement at the upper extremes of the score scale.

In the case of writing, it may be that scorers, given a limited amount of training and a great many papers to score, were unable to internalize and

Table 6

Percent Agreement Comparison of Writing Scores From First and Second Scorings for Memories (#18) (*N*=439)

First score	Second score 0	1	2	3	4	5	6	Total %	*N*
0	50.00	0.00	31.25	12.50	6.25	0.00	0.00	100.00	16
1	3.80	55.70	35.44	3.80	1.27	0.00	0.00	100.00	79
2	1.02	9.18	58.67	22.96	7.14	1.02	0.00	100.00	196
3	2.22	7.78	32.22	36.67	21.11	0.00	0.00	100.00	90
4	0.00	6.82	22.73	38.64	25.00	4.55	2.27	100.00	44
5	0.00	0.00	18.18	36.36	27.27	18.18	0.00	100.00	11
6	0.00	0.00	0.00	0.00	66.67	0.00	33.00	100.00	3
N	15	72	189	104	51	6	2		

Table 7

Percent Agreement Comparison of Writing Scores From First and Second Scorings for Camels (#4) (*N*=237)

First score	Second score 0	1	2	3	4	5	6	Total %	*N*
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0
1	0.00	68.75	25.00	3.12	3.12	0.00	0.00	100.00	32
2	0.00	10.81	45.94	33.78	9.46	0.00	0.00	100.00	74
3	0.00	2.33	40.70	36.05	16.28	4.65	0.00	100.00	86
4	0.00	0.00	19.35	32.26	35.48	9.68	3.23	100.00	31
5	0.00	0.00	10.00	40.00	30.00	20.00	0.00	100.00	10
6	0.00	0.00	0.00	50.00	0.00	50.00	0.00	100.00	4
N	0	32	84	73	36	11	1		

Table 8

Percent Agreement Comparison of Writing Scores From First and Second Scorings for Wolves (#3) (*N*=705)

First score	Second score							Total %	*N*
	0	1	2	3	4	5	6		
0	50.00	43.75	6.25	0.00	0.00	0.00	0.00	100.00	26
1	3.80	59.50	35.44	1.27	0.00	0.00	0.00	100.00	127
2	0.00	9.18	59.69	11.96	19.16	0.00	0.00	100.00	314
3	0.00	10.00	34.33	34.56	21.11	0.00	0.00	100.00	144
4	0.00	6.82	22.73	38.64	25.00	4.55	2.27	100.00	71
5	0.00	0.00	15.18	34.36	32.28	18.18	0.00	100.00	18
6	0.00	0.00	0.00	0.00	0.00	60.00	40.00	100.00	5
N	18	135	302	122	114	11	3		

apply criteria for all levels of a 7-point scale. It may also be that the tasks did not stimulate student responses that spanned the full scale. In any case, as in reading, these findings suggest that to increase interscorer reliability, the score scale must reflect actual variations in student responses to tasks, rubrics and exemplars should be selected to clearly differentiate between score levels, and training must be intensified.

Math

Table 9 reports indicators of interscorer comparability for the seven math tasks scored at the Big Sky Scoring Conference.

Correlation. Interscorer correlations for holistic scoring of the seven math tasks were higher on the average than holistic scores for reading and writing. In fact, for six out of the seven math tasks, The Aquarium Problem (*r* = .76), Name that Graph (*r* = .69), Hot Dog Rolls, (*r* = .75), Pizza Party (*r* = .72), Building with Tiles (*r* = .66), and Too High, Too Low (*r* = .78), correlations between scores assigned by two different raters approached or exceeded acceptable levels (>.75). Only the Amusement Park task (#8) produced a low interscorer reliability coefficient.

Table 9

Indicators of Interscorer Comparability for Seven Math Tasks

Task	Correlation	% Agreement	% Agreement (±1 point)
The Aquarium Problem (#73) (*N*=141)	.76	47%	91%
Amusement Park (#8) (*N*=143)	.44	40%	86%
Name that Graph (#9) (*N*=64)	.69	48%	87%
Hot Dog Rolls (#10) (*N*=183)	.75	58%	94%
Pizza Party (#12) (*N*=130)	.72	44%	91%
Building with Tiles (#16) (*N*=148)	.66	55%	90%
Too High, Too Low (#21) (*N*=147)	.78	50%	95%

Percent agreement. Despite the higher correlation coefficients, exact percent agreement for math tasks averaged 49%, comparable to exact agreement on the literacy tasks. Adjacent percent agreement on math tasks was very high, averaging 91% across the seven tasks. If we look at the bivariate distributions comparing first score with second score for the seven math tasks (Table 10 through Table 16), once again we see that scorers were able to agree most frequently on papers at the low end of the score range (0s, 1s, and 2s on a 7-point scale). They most often disagreed on midrange papers (3s, 4s, and 5s on a 7-point scale). There were too few student responses scored "6" in the double-scored sample to assess agreement at the upper extreme of the score scale.

Table 10

Percent Agreement Comparison of Scores From First and Second Scorings for the Aquarium Problem (*N*=115)

First score	Second score 0	1	2	3	4	5	6	Total %	*N*
0	50.00	50.00	0.00	0.00	0.00	0.00	0.00	100.00	2
1	0.00	67.67	33.33	0.00	0.00	0.00	0.00	100.00	6
2	0.00	21.43	35.71	21.43	21.43	0.00	0.00	100.00	28
3	0.00	0.00	32.26	45.16	19.35	3.23	0.00	100.00	31
4	0.00	0.00	9.38	9.38	53.13	28.13	0.00	100.00	32
5	0.00	0.00	0.00	0.00	50.00	42.86	7.14	100.00	14
6	0.00	0.00	0.00	0.00	0.00	0.00	100.00	100.00	2
N	1	11	25	23	36	16	3		

Table 11

Percent Agreement Comparison of Scores From First and Second Rounds of Scoring for The Amusement Park (*N*=135)

First score	Second score 1	2	3	4	5	Total %	*N*
1	57.14	42.86	0.00	0.00	0.00	100.00	7
2	12.12	42.42	33.33	9.09	3.03	100.00	28
3	0.00	18.87	45.28	28.30	7.55	100.00	53
4	0.00	7.14	46.43	28.57	17.86	100.00	28
5	0.00	5.88	35.29	11.76	47.06	100.00	17
N	4	29	55	28	19		

Table 12

Percent Agreement Comparison of Scores From the First and Second Scorings for Name that Graph (*N*=69)

First score	Second score 0	1	2	3	4	5	6	Total %	*N*
0	100.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	3
1	0.00	40.00	40.00	20.00	0.00	0.00	0.00	100.00	5
2	0.00	14.29	71.43	14.29	0.00	0.00	0.00	100.00	7
3	0.00	0.00	24.14	44.83	27.59	3.45	0.00	100.00	29
4	0.00	0.00	11.76	23.53	47.06	17.64	0.00	100.00	17
5	0.00	0.00	14.29	28.57	28.57	28.57	0.00	100.00	7
6	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00	1
N	3	3	17	21	18	7			

Table 13

Percent Agreement Comparison of Scores From First and Second Scorings for Hot Dog Rolls (*N*=148)

First score	Second score 0	1	2	3	4	5	6	Total %	*N*
0	100.00	00.00	0.00	0.00	0.00	0.00	0.00	100.00	2
1	0.00	75.00	16.67	8.33	0.00	0.00	0.00	100.00	12
2	0.00	21.74	52.17	26.09	0.00	0.00	0.00	100.00	23
3	0.00	4.88	24.39	46.34	21.95	2.44	0.00	100.00	41
4	2.78	0.00	5.56	27.78	58.33	5.56	0.00	100.00	36
5	3.03	0.00	0.00	6.06	15.15	72.72	3.03	100.00	33
6	0.00	0.00	0.00	0.00	0.00	100.00	0.00	100.00	1
N	3	16	26	39	35	28	1		

Table 14

Percent Agreement Comparison of Scores From First and Second Scorings for Pizza Party (*N*=117)

First score	Second score							Total %	*N*
	0	1	2	3	4	5	6		
0	100.00	00.00	0.00	0.00	0.00	0.00	0.00	100.00	1
1	7.69	76.92	0.00	7.69	7.69	0.00	0.00	100.00	13
2	0.00	31.58	42.11	21.05	5.26	0.00	0.00	100.00	19
3	0.00	0.00	35.29	32.35	26.47	5.88	0.00	100.00	34
4	0.00	6.45	0.00	35.48	35.48	22.58	0.00	100.00	31
5	0.00	5.56	0.00	11.11	22.22	50.00	11.11	100.00	18
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	1
N	2	19	20	29	26	18	3		

Table 15

Percent Agreement Comparison of Scores From First and Second Scorings for Building with Tiles (*N*=117)

First score	Second score							Total %	*N*
	0	1	2	3	4	5	6		
0	100.00	00.00	0.00	0.00	0.00	0.00	0.00	100.00	2
1	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100.00	10
2	18.75	6.25	25.00	50.00	0.00	0.00	0.00	100.00	16
3	6.98	0.00	13.95	60.47	18.60	0.00	0.00	100.00	43
4	3.23	0.00	3.23	25.81	41.94	25.81	0.00	100.00	31
5	0.00	0.00	0.00	6.67	13.33	80.00	0.00	100.00	15
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0
N	7	10	12	44	23	21	0		

Table 16

Percent Agreement Comparison of Scores From First and Second Scorings for Too High, Too Low (*N*=142)

First score	Second score 0	1	2	3	4	5	6	Total %	*N*
0	100.00	00.00	0.00	0.00	0.00	0.00	0.00	100.00	4
1	0.00	58.82	25.29	5.88	0.00	0.00	0.00	100.00	17
2	0.00	15.79	50.00	31.58	2.63	0.00	0.00	100.00	38
3	0.00	2.27	29.55	52.27	13.64	2.27	0.00	100.00	44
4	0.00	0.00	3.57	60.71	28.57	3.57	3.57	100.00	28
5	0.00	0.00	0.00	9.10	18.18	63.63	9.10	100.00	11
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0
N	4	17	39	54	17	9	2		

Comparability Across Scoring Methods for Reading

In reading, about 50% of the student responses for two tasks, Folk Tales (#17) and Camels (#4), were scored holistically and anaholistically. Correlations between scores assigned using holistic scoring methods and those assigned using anaholistic methods were computed as indicators of the comparability of scoring methods. These findings are reported on Tables 17 and 18. The low interscorer reliabilities associated with both holistic and anaholistic scoring procedures used at Big Sky limit the meaning of any analysis of the resulting scores. These and other findings should be interpreted with extreme caution.

Across both tasks, there was a moderate relationship between the holistic score and each item score (Range: .43 to .64). The correlation between the holistic score and the "Across" score was moderate (*r* = .61) for Folk Tales and low (*r* = .26) for Camels. The most consistently strong relationship was between the holistic score and the "Sum" score produced by adding individual item scores (*r* = .67 for Folk Tales and .63 for Camels). This finding, coupled with the higher cross-scorer comparability of the sum score reported above, favors the use of the sum rather than individual item or "Across" scores in anaholistic scoring as a composite measure.

Table 17

Correlation Matrix. Indicators of Cross-method Comparability of Holistic and Anaholistic Reading Scoring for Folk Tales (#17) (*N*=313)

	Anaholistic scores								
	Item 1	Item 2	Item 3	Item 4	Item 5–6	Item 7	Item 8	Across item	Sum
Holistic score	.46	.65	.59	.56	.59	.45	.43	.61	.67

Table 18

Correlation Matrix. Indicators of Cross-method Comparability of Holistic and Anaholistic Reading Scoring for Camels (#4) (*N*=513)

	Anaholistic scores						
	Item 1	Item 2	Item 3	Item 4	Item 5	Across item	Sum
Holistic score	.64	.50	.51	.48	.50	.26	.63

Distributions of Student Performance

Given the unsystematic selection of the student sample for the 1992 spring pilot, summaries of student performance are difficult to interpret. The means, standard deviations, frequencies, and percentage frequency distributions reported for all tasks in Table 19 through Table 23 are intended to provide descriptive information about the tasks for purposes of task improvement and rubric refinement. Use of these data to make judgments about the quality of students' performance or educational programs is not appropriate.

Across all tasks in all content areas, comparisons of score ranges, means, standard deviations, and frequency distributions showed a remarkable consistency in judgments of students' performance. For virtually all tasks, distribution of judgments of student performance were unimodal and positively skewed. Few student responses were assigned the highest score ratings. Modal ratings were in the low to midrange of the score scale.

Table 19

Means, Standard Deviations, Frequencies and Percentage Frequency Distributions of Holistic Scores for All Reading Tasks

Task	Score				
	0	1	2	3	4
Folk Tales (#17) (*M*=2.00; *SD*=.85; *N*=777)					
Frequency	24	189	351	191	22
Percent	3.09	24.32	45.17	24.58	2.83
Camels (#4) (*M*=1.83; *SD*=.87; *N*=642)					
Frequency	20	217	281	98	26
Percent	3.12	33.80	43.77	15.26	4.05
Wolves (#3) (*M*=1.98; *SD*=.74; *N*=759)					
Frequency	6	180	412	146	15
Percent	0.79	23.68	54.21	19.21	1.97

Table 20

Means, Standard Deviations, Frequencies and Percentage Frequency Distributions of Reading Scores for Anaholistically Scored Camels (#4) (*N*=728)

Item	Score				
	0	1	2	3	4
Item 1 (*M*=2.12, *SD*=.86)					
Frequency	17	133	372	160	46
Percent	2.33	18.24	51.03	21.95	6.31
Item 2 (*M*=1.88; *SD*=.84)					
Frequency	31	190	367	116	24
Percent	4.25	26.06	50.34	15.91	3.29
Item 3 (*M*=1.91; *SD*=.76)					
Frequency	22	162	424	101	19
Percent	3.02	22.22	58.16	13.85	2.61
Item 4 (*M*=1.59; *SD*=.84)					
Frequency	54	297	281	88	8
Percent	7.41	40.74	38.55	12.07	1.10
Item 5 (*M*=1.67; *SD*=.87)					
Frequency	63	235	317	102	11
Percent	8.64	32.24	43.48	13.99	1.51
Across items (*M*=1.16; *SD*=1.03)					
Frequency	261	158	245	60	4
Percent	35.80	21.67	33.61	8.23	0.55

Table 21

Means, Standard Deviations, Frequencies and Percentage Frequency Distributions of Reading Scores for Anaholistically Scored Folk Tales (#17) (*N*=247)

Item	Score				
	0	1	2	3	4
Item 1 (*M*=2.08; *SD*=.82)					
Frequency	3	57	112	67	8
Percent	1.21	23.08	45.34	27.13	3.24
Item 2 (*M*=2.27; *SD*=.78)					
Frequency	6	23	129	76	13
Percent	2.43	9.31	52.23	30.77	5.26
Item 3 (*M*=2.08; *SD*=.91)					
Frequency	15	38	118	65	11
Percent	6.07	15.38	47.77	26.32	4.45
Item 4 (*M*=1.92; *SD*=.96)					
Frequency	19	56	111	49	12
Percent	7.69	22.67	44.94	19.84	4.86
Item 5–6 (*M*=1.94; *SD*=1.69)					
Frequency	19	57	100	61	10
Percent	7.69	23.08	40.49	24.70	4.05
Item 7 (*M*=1.69; *SD*=.99)					
Frequency	35	59	105	43	5
Percent	14.17	23.89	42.51	17.41	2.02
Item 8 (*M*=1.56; *SD*=1.12)					
Frequency	53	65	74	47	8
Percent	21.46	26.32	29.96	19.03	3.24
Across items (*M*=2.03; *SD*=.75)					
Frequency	5	45	140	51	6
Percent	2.02	18.22	56.68	20.65	2.43

Table 22

Means, Standard Deviations, Frequencies and Percentage Frequency Distribution of Scores for All Writing Tasks

Task	Score						
	0	1	2	3	4	5	6
Memories (#18) (*M*=2.24; *SD*=1.14; *N*=957)							
Frequency	61	145	413	211	100	19	8
Percent	6.37	15.15	43.16	22.05	10.45	1.99	0.84
Camels (#4) (*M*=2.47; *SD*=1.21; *N*=325)							
Frequency	18	43	109	98	41	12	4
Percent	5.54	13.23	33.54	30.15	12.62	3.69	1.23
Wolves (#3) (*M*=2.21; *SD*=1.15; *N*=924)							
Frequency	55	138	435	187	85	15	9
Percent	5.95	14.90	47.00	20.23	9.19	1.62	0.97

Table 23

Means, Standard Deviations, Frequencies and Percentage Frequency Distributions of Scores for All Math Tasks

Task	Score						
	0	1	2	3	4	5	6
Aquarium Problem (#73) (*M*=3.17; *SD*=1.28; *N*=700)							
Frequency	16	55	143	176	204	100	7
Percent	2.29	7.86	20.43	25.14	29.14	14.29	0.97
The Amusement Park (#8) (*M*=3.19; *SD*=1.15; *N*=886)							
Frequency	13	46	172	309	218	128	0
Percent	1.50	5.20	19.40	34.90	24.60	14.40	0
Name that Graph (#9) (*M*=2.79; *SD*=1.21; *N*=502)							
Frequency	20	55	108	183	106	25	25
Percent	4.00	11.00	21.50	36.40	21.10	5.00	1.00
Hot Dog Rolls (#10) (*M*=3.22; *SD*=1.32; *N*=793)							
Frequency	14	81	136	209	193	158	2
Percent	1.80	10.20	17.20	26.40	24.30	19.90	0.30
Pizza Party (#12) (*M*=3.14; *SD*=1.33; *N*=630)							
Frequency	8	74	114	182	147	90	15
Percent	1.30	11.70	18.10	28.90	23.30	14.30	2.40
Building with Tiles (#15) (*M*=3.29; *SD*=1.11; *N*=648)							
Frequency	6	35	93	240	183	86	5
Percent	0.90	5.40	14.40	37.00	28.00	13.30	0.80
Too High, Too Low (#21) (*M*=2.87; *SD*=1.27; *N*=765)							
Frequency	14	90	213	210	147	83	8
Percent	1.80	11.80	27.80	27.50	19.20	10.80	1.00

SUMMARY

The major findings of the analysis of the 1992 spring pilot, reported in detail above, are summarized below:

- **Interscorer reliability estimates for reading and writing tasks were in the moderate range, below levels achieved with the use of large-scale writing assessment or standardized tests. Low reliability limits the use of 1992 reading and writing scores for making judgments about student performance or educational programs.**

- **Interscorer reliability estimates for math tasks were somewhat higher than for literacy. For six out of seven math tasks, reliability coefficients approached or exceeded acceptable levels.**

- **Differences existed between exact scores assigned to a student response by different scorers, but the scores seldom differed by more than one point, even on 7-point scales.**

- **Cross-scorer differences were most pronounced in midrange papers.**

- **Use of anaholistic and holistic scoring methods resulted in different scores for the same student response. The anaholistic "Sum" score was most similar to the holistic rating.**

- **Distributions of scores were unimodal and positively skewed for virtually all tasks. Tasks within a content area had similar means and standard deviations.**

CONCLUSIONS AND IMPLICATIONS

Sampling limitations and low interscorer reliability preclude the use of 1992 spring pilot data for making judgments about the quality of student performance or educational programs, and limit many analyses that would otherwise be carried out to validate the assessment results. However, analysis of 1992 spring pilot data does provide some direction for refinement and revision of tasks, rubrics, and scorer training. With regard to reliability, attention should be paid to factors that contribute to lack of agreement among raters such as insufficient training or ambiguous or overly complex rubrics and extended score scales.

Training

The nature of training at Big Sky was not evaluated in any systematic way, other than asking participants how they felt about the experience and examining the quality of the data produced. Participants at the Big Sky Scoring Conference spoke of feeling overwhelmed and overworked after four days of learning to score and scoring. Especially for literacy tasks, scorers were asked to internalize and apply lots of information very quickly. Low interscorer reliability indicated that scorers had difficulty applying rubrics and assigning scores in a consistent manner.

The strategy at Big Sky was, as part of the professional development component of the New Standards Project, to expose a large number of teachers and other professionals from the partners to mass scoring of on-demand tasks. This may have been warranted for professional development and public engagement purposes; however, the large number and varied nature of participants may have jeopardized the production of valid and reliable information about student performance. To increase the reliability of scoring, it would be advantageous to provide intensive training for a small, homogeneous group of scorers in any given year. The training should include evaluation procedures to assure that trained scorers have reached a high level of proficiency for a task before they are permitted to score that task independently.

Training should encourage scorers to become familiar with the rubrics and the range of student performance for particular tasks. Group scoring, think alouds, and careful review of benchmark papers are strategies that have been used to increase reliability of scores. All of these activities take time. If they are to occur, the amount of time allotted for the training will need to be increased substantially. This may mean that one conference does not allow sufficient time to train and score, or that a single scorer cannot be trained to score, multiple tasks. Partners may be better off with individual scorers who are trained to score a single task, while the capacity to score multiple tasks rests with the team as a whole.

Rubrics

Aspects of the scoring systems also may have contributed to low reliabilities on some tasks. Unclear or inconsistent language in the rubrics or scoring directions can be interpreted differently by different scorers and is often a source of error in double-scored papers. For example, for literacy and math tasks, bivariate distributions showed that scorers had difficulty making distinctions between midrange values of the score scale. Review of the rubrics at these values for the reading tasks shows the similarity between descriptions of performance at level "3" and level "4" on a 0 to 4 scale:

> **Level 3: Responses demonstrate an adequate understanding of the text. There is evidence to understanding of both the gist and specific parts of the text. Level three responses are more complex than responses at levels one or two. They may include minimal extensions, such as connection to other texts, experiences, abstractions and/or generalization. All elements of the question are addressed in the response.**
>
> **Level 4: Responses are complex and demonstrate a thorough understanding and interpretation of the text. There is considerable evidence of extension of the text, such as connection to other texts, experiences, abstractions and/or generalization. There may be evidence of "reading like a writer"—attending to, evaluating, or appreciating the author's perspective and craft in creating a text. All elements of the question are addressed in the response.**

Common understanding of words like "complex," and "considerable" versus "minimal" would seem essential for scorers to be able to reliably differentiate between levels 3 and 4 when judging students' work. Exemplars, benchmark training, and explicit definitions of key terms enhance common understanding and interscorer agreement. In all cases, rubrics should stress the factors that differentiate each level from adjacent levels and training should teach scorers to recognize the distinctions through the use of well-chosen exemplars and clear language used to define levels.

Even after rubrics are perfected and scoring is refined, there may be instances where agreement is not reached for specific score levels. In this case, one might consider simplifying the scoring system to include fewer levels.

Comparison of Alternate Scoring Methods

Even with limitations imposed by low reliabilities, comparison of holistic and anaholistic methods gave some insight into the comparative adequacy and utility of the resulting scores. Additional comparative analyses of this type can help in the design and selection of scoring systems and should be built into future scoring conferences.

Despite sampling limitations and low reliabilities that limited the use of the 1992 spring pilot data, results of the Big Sky Scoring Conference provided evidence that scoring of large-scale performance assessment can be done and done well, but it requires ample time for training, evaluation, feedback and discussion, clear definitions of levels and the distinctions among them, sufficient, well-chosen exemplars, and lots of hard work!

References

Gronlund, N.E., & Linn, R.L. (1990). ***Measurement and evaluation in teaching*** **(6th ed.). New York: Macmillan.**

Vermont Department of Education (1991). ***"This is my best": Vermont's Writing Assessment Program, pilot year 1990-1991.*** **Montpelier: Author.**