DOCUMENT RESUME

ED 427 080 TM 029 462

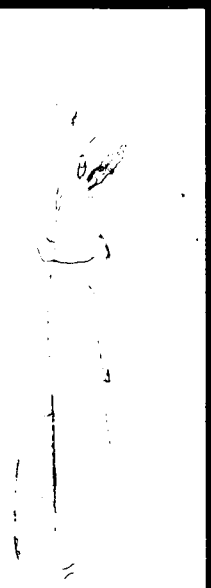
AUTHOR Waltman, Kristie; Kahn, Andrea; Koency, Gina
TITLE Alternative Approaches to Scoring: The Effects of Using Different Scoring Methods on the Validity of Scores from a Performance Assessment. CSE Technical Report 488.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-488
PUB DATE 1998-10-00
NOTE 43p.
CONTRACT R305B60002
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Holistic Approach; Intermediate Grades; Junior High Schools; Middle Schools; *Performance Based Assessment; *Science Tests; *Scoring; Training; *Validity
IDENTIFIERS Scoring Rubrics

ABSTRACT

The degree to which modifications to the format of the scoring rubric and the associated training procedures affect the technical quality of the resulting scores was studied, and the perceived utility of each scoring method for influencing a teacher's instructional decisions positively was investigated. Two different methods were used to score responses to six middle-school science performance tasks completed by 100 to 200 students. Although both types of scoring could be characterized as focused holistic, the format and training associated with how the scoring criteria are presented and used by rates were modified to create two different methods: focused holistic (F-H) and analytic impression (A-I). Evidence form the study suggests that the F-H and A-I methods are not equally preferable for making decisions about individual students or groups of students. However, the rates were overwhelmingly in favor of the A-I method for obtaining useful information to improve instruction. Appendixes contain the scoring rubrics for problem-solving and explanation performance tasks. (Contains 10 tables, 5 figures, and 5 references.) (Author/SLD)

CRESST

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

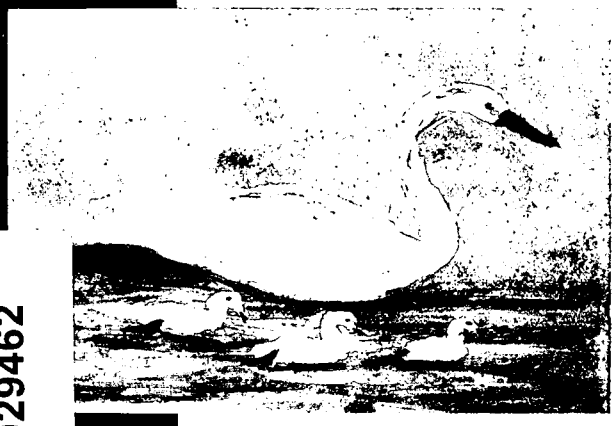


Alternative Approaches to Scoring: The Effects of Using Different Scoring Methods on the Validity of Scores From a Performance Assessment

CSE Technical Report 488

Kristie Waltman
American College Testing

Andrea Kahn and Gina Koency
CRESST/University of California, Los Angeles



UCLA Center for the Study of Evaluation

In Collaboration With:

University of Colorado at Boulder • Stanford University • The RAND Corporation
University of California, Santa Barbara • University of Southern California
Educational Testing Service • University of Pittsburgh

Alternative Approaches to Scoring: The Effects of Using Different Scoring Methods on the Validity of Scores From a Performance Assessment

CSE Technical Report 488

Kristie Waltman
American College Testing

Andrea Kahn and Gina Koency
CRESST/University of California, Los Angeles

October 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.2 Assessment in Action: Making a Difference in Practice—Issues in System Coherence.
Joan L. Herman, Project Director, CRESST

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, 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 positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

ALTERNATIVE APPROACHES TO SCORING: THE EFFECTS OF USING DIFFERENT SCORING METHODS ON THE VALIDITY OF SCORES FROM A PERFORMANCE ASSESSMENT[1]

Kris Waltman
American College Testing

Andrea Kahn and Gina Koency
CRESST/University of California, Los Angeles

Abstract

The purpose of this study was to investigate the degree to which modifications to the format of the scoring rubric and the associated training procedures affect the technical quality of the resulting scores and the perceived utility of each scoring method for positively influencing a teacher's instructional decisions. Two different methods were used to score responses to six middle-school science performance tasks. Although both types of scoring utilized in this study could be characterized as focused holistic, the format and training associated with how the scoring criteria are presented and utilized by raters were modified to create two different methods—Focused Holistic (F-H) and Analytic Impression (A-I). Evidence from the study suggests that the F-H and A-I methods are not equally preferable for making decisions about individual students or groups of students. However, the raters were overwhelmingly in favor of the A-I method for obtaining useful information to improve instruction.

Introduction

The use of performance tasks to evaluate student achievement is encumbered by both practical constraints and threats to technical quality. Although some have questioned the need for performance assessments to be held to the same technical rigor as more traditional measures (Gipps, 1994; Moss, 1994), in the context of high-stakes accountability it is preferable that the scores resulting from these assessments be as high in technical quality as possible given the context of the testing situation. The desire is to obtain reliable scores in as efficient a manner as possible that will lead to valid interpretations. This goal, however, typically is not fully realized due to the practical constraints of limited

[1] For simplicity, the term *method* is used in this paper to refer to both the format of the scoring rubric and the associated training procedures.

resources available for scoring (e.g., time and money). The purpose of this study was to obtain evidence that can be used to assist decision makers in obtaining scores that have optimal technical quality given the context and purpose of the assessment.

It has been recognized that one way in which rater consistency (both inter and intra) can be increased is by using clearly delineated scoring criteria that are rigidly defined (Wainer, 1993). When the same criteria are used to evaluate performance on different tasks (sometimes referred to as a *generalized* rubric) it is possible that having criteria too highly specified will result in (a) decreased rater consistency on a given task, (b) differential application of the criteria across tasks (contributing to a lack of generalizability across tasks), and (c) increased time needed for both training and scoring. In addition to the specificity of the scoring criteria, the manner in which raters are trained also plays an integral role in determining consistency within the scoring process (Herman, Aschbacher, & Winters, 1992). Often, however, the practical considerations of time and money determine the amount and degree of training that can be utilized. This trade-off between technical quality and practical constraints is a dilemma too often faced by those responsible for developing the procedures associated with scoring performance assessments.

Although it is recognized that elements of the scoring process influence rater consistency, under a generalizability framework both the scoring rubric and the training procedures are typically considered to be a fixed facet rather than a random facet (Brennan, 1996). The facets are considered *fixed* in that only a single rubric and set of training procedures are used, and all inferences made based on the scores are conditioned on the use of the particular rubric and procedures. However, Brennan notes that, in principle, a given performance task could be scored by any number of rubrics (and raters could be trained to use them in different ways), and he uses this situation to provide an example of the "blurred distinction between reliability and validity." According to Brennan, "If two or more rubrics are in principle equally acceptable, then the issue is primarily in the realm of reliability. However, if the acceptable rubrics are not equally preferable, then the matter is largely one of validity." This study was designed to investigate whether modifications in the way the scoring criteria are presented to the raters and in the way in which raters are trained to use the scoring criteria will result in

making one scoring method more preferable than another, and thus in raising questions about validity.

Scoring Methods

Although both types of scoring to be utilized in this study could be characterized as focused holistic (i.e., a single score is provided using absolute criteria that are associated with several dimensions), the format and training associated with how these criteria are presented and utilized by the raters were modified to create two different scoring methods—Focused Holistic (F-H) and Analytic Impression (A-I). The F-H format summarizes the criteria collectively within a score point and does not emphasize the dimensional aspect of each criterion. In contrast, the A-I format separates out these same criteria for each score point across the dimensions, much like what is needed for typical analytic scoring methods. Unlike analytic scoring, however, the purpose of the A-I method is not to yield a separate score for each of the dimensions, but rather to use the separate dimensions to better understand the meaning associated with each of the score points and determine which overall score "best fits" the student's performance. (An example of the F-H and A-I formats of the scoring rubrics can be found in Appendix A for Problem Solving and Appendix B for Explanation.) The training procedures associated with each of these two rubric formats also differ. The emphasis of the training associated with the F-H scoring rubric was on identifying the set of criteria associated with a particular score point that best captures the student performance, whereas the training emphasis for the A-I scoring rubric was on evaluating performance with respect to each dimension separately and using the dimensional judgments to generate a single score.

In training raters to use the F-H scoring method, raters were first presented with a description of the dimensions embedded in each score point. Raters were instructed to evaluate each set of criteria as a whole and to determine the set of criteria that best reflects the student performance. To facilitate the process, raters were asked to quickly identify one or two score points that appeared to be most appropriate. Then, they were instructed to analyze carefully the criteria associated with each score point to determine the one that best characterizes the response. During training it was emphasized that a response did not need to match all the criteria associated with a given score point, but that overall the chosen set of criteria should be a better description of the response than any other set of

criteria. Thus when using the F-H scoring rubric, raters were not instructed to evaluate each dimension separately to generate an overall score. Rather, raters were trained to identify the set of criteria that best captured the student performance across the dimensions.

Conversely, when training raters to use the A-I scoring method, emphasis was placed on evaluating performance on each dimension separately and then assigning a single score based on the dimensional judgments. In this way, there was not a fixed set of criteria associated with each score point. Raters could arrive at a given score point in numerous ways. To facilitate the scoring process, raters were presented with different approaches. For example, raters could indicate the level of performance on each dimension by placing an adhesive marker (tape flag) on the appropriate criterion for each dimension. Then they were instructed to analyze the distribution of the markers across the dimensions to determine the score point that overall "best fit" the response. Another approach simply called for raters to record scores for each dimension and used the score values to generate a final, overall score. Raters were free to adopt their preferred approach when scoring the problem-solving tasks, but they were asked to record their dimensional scores for the explanation tasks. It is important to note that raters were not given explicit guidelines on how to combine information across dimensions. Instead, they were instructed to use their professional judgment in determining how much weight should be given to each dimension. Moreover, although dimensional judgments were discussed during training, rater agreement on each dimension was not explicitly evaluated. Rater agreement was emphasized only with respect to the overall score.

Purpose

The purpose of this study was to investigate the degree to which modifications to the format of the scoring rubric and the associated training procedures affect the technical quality of the resulting scores and the perceived utility of each scoring method for positively influencing a teacher's instructional decisions. The specific research questions investigated in this study were:

1. How do the distributions of scores resulting from the F-H and A-I methods compare?
2. How do the interrater consistency indices resulting from the use of the F-H and A-I methods compare?

3. How do the correlations between the scores resulting from the F-H and A-I methods and scores from a related measure of a student's prior knowledge compare?
4. How do the ratings provided for each of the dimensions compare to the overall A-I score for the explanation performance tasks?
5. How do the raters' perceptions differ regarding the ease of using the scoring process associated with the F-H and A-I methods?
6. How do the raters' perceptions differ regarding the scoring method that would have the greatest utility for positively influencing their instructional decisions?

Method

Instruments

The performance tasks used for this study consisted of six middle-school science tasks that were administered to approximately 100 to 200 seventh- or eighth-grade students in an urban school district as part of a large-scale pilot test for a test development effort. Students did not respond to all six performance tasks; rather, different combinations of performance tasks were randomly assigned to each of 10 classrooms. The three content areas measured by these performance tasks were ecosystems (life science), land forms (Earth science), and density (physical science). Each of these three content areas was represented by two different types of performance tasks—problem solving and explanation. These task types require different scoring criteria, and thus, two sets of scoring rubrics were used. (The scoring rubrics can be found in Appendices A and B, respectively, for problem solving and explanation.) The names of each of the specific performance tasks, by content area, are provided in Table 1. Students were also administered an associated prior knowledge test consisting of 36 multiple-choice items, with 12 questions for each of the three content areas.

Table 1

Names of Performance Tasks by Content Area

	Performance task	
Content area	Problem solving	Explanation
Life Science	Killer Bees	Yellowstone
Earth Science	Mountain Hike	Land Over Time
Physical Science	Density and Matter	Density

Experimental Design and Scoring Process

Given the purpose of this study, a counterbalanced design was utilized in order to look solely at the effects due to scoring methods. Effects due to different groups of raters and different type of performance task were confounded. The design, illustrated in Figure 1, required two groups of raters (Group A and Group B) to score one type of performance tasks with the F-H method and to score the remaining type of performance tasks using the A-I method. Thus, four separate scoring sessions were conducted.

In order to minimize the lasting effects of the training procedures associated with each group's first scoring session, the scoring sessions were scheduled so that approximately three months separated the first and second meetings for each group. The problem-solving tasks were scored during the first two-day meeting, and the explanation tasks were scored during the second. The teachers were randomly divided into two groups, Group A and Group B, prior to the first meeting. During the first meeting, when the three problem-solving tasks were scored, Group A was trained using the F-H method and Group B was trained using the A-I method. Three months later, when the three explanation tasks were scores, Group A was trained using the A-I method, and Group B was trained using the F-H method. Thus, each response was scored four times—by two raters in Group A and by two raters in Group B. It should be noted, however, that although each response was scored by two raters the design is still unbalanced because all raters did not score all responses.

The raters participating in the scoring sessions were practicing middle-school science teachers; thus, they all had experience at the grade level of the students whose performance they were to evaluate. Although a total of 16 teachers had originally agreed to participate in the study, several teachers were

Performance Tasks	Scoring Method	
	F-H	A-I
Problem Solving	1) Group A	2) Group B
Explanation	3) Group B	4) Group A

Figure 1. Experimental design for the scoring sessions.

lost due to attrition (primarily due to scheduling conflicts). Table 2 summarizes the number of raters participating on each day of the scoring sessions. Due to the large disparity in the number of raters for the explanation tasks, the complete set of student responses to these performance tasks could not be scored using both methods. More specifically, Group B was able to score only 65% of the total set of student responses scored by Group A.

Two scores were generated for each response to the performance tasks—one each for the F-H and A-I methods—by summing the two ratings (from the 0- to 5-point scale) obtained from within each method. Thus, the score scale for these summed ratings ranges from zero to 10. In addition, within each scoring method scores for three "combined performance tasks" were computed—Life Science, Earth Science, and Problem Solving. These scores were calculated in the following manner: (a) Life Science = Killer Bees + Yellowstone; (b) Earth Science = Mountain Hike + Land Over Time; and (c) Problem Solving = Killer Bees + Mountain Hike. Due to the extreme level of difficulty associated with the two physical science tasks, neither Density and Matter nor Density was used to compute composite scores. Furthermore, an Explanation composite score could not be formed due to the small number of students who responded to both Yellowstone and Land Over Time. The score scales for each of the three combined sets of performance tasks ranges from zero to 20.

Data Analysis

The distributions of scores obtained from the two scoring methods were compared, separately for each task and for the combined sets of performance tasks, using paired *t*-tests. Because no one interrater consistency index can capture the entire picture of how two sets of scores are related, several indices were calculated in order to better understand the agreement of raters within and

Table 2

Number of Teachers Participating in Each Scoring Session

	F-H		A-I	
Scoring session	Day 1	Day 2	Day 1	Day 2
Problem solving	8	8	8	7
Explanation	5	3	7	7

between the two scoring methods. First, within each scoring method, the two ratings provided for each student response were used to calculate the percent of exact agreement and the percent of agreement within one score point. In addition, Kappa was estimated to determine the degree to which rater agreement was beyond that which could be attributable to chance agreement, and the product-moment correlation between the two sets of ratings was calculated for each performance task.

Each of the interrater consistency indices was also calculated to estimate the agreement of scores across the two scoring methods. These values were intended to help answer the question "How do ratings from the two scoring methods compare for an individual student's response?" Interrater consistency between the two methods was calculated in two ways. First, the methods were compared by calculating the consistency of the first scores that were provided within each scoring session. This comparison is represented by "First" in Figure 2. The second comparison involved the summed scores from each method, represented by "Summed" in Figure 2.

Although no criterion measure was available to obtain criterion-related validity evidence for scores from the performance tasks, students were administered a test consisting of 36 multiple-choice items that was designed to provide evidence of the students' prior knowledge in the associated content areas. Bivariate correlations were computed between the scores on the prior knowledge test and each set of scores for the performance tasks in order to determine the degree of linear relationship between them and to determine

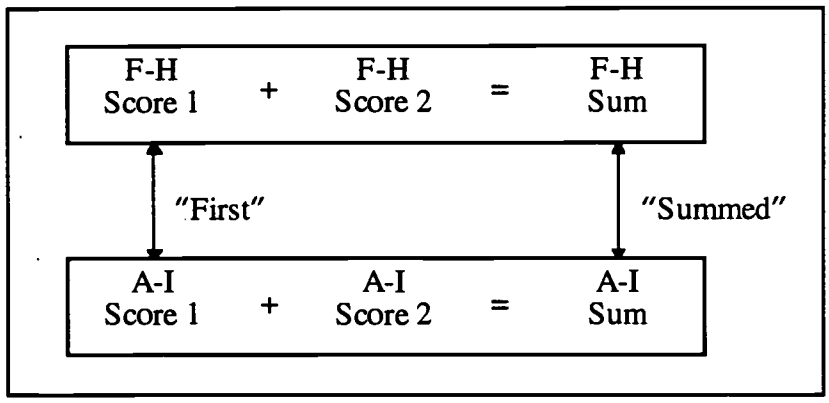


Figure 2. Comparisons of interrater agreement across methods.

whether scores from one of the two scoring methods were more highly correlated with the prior knowledge test.

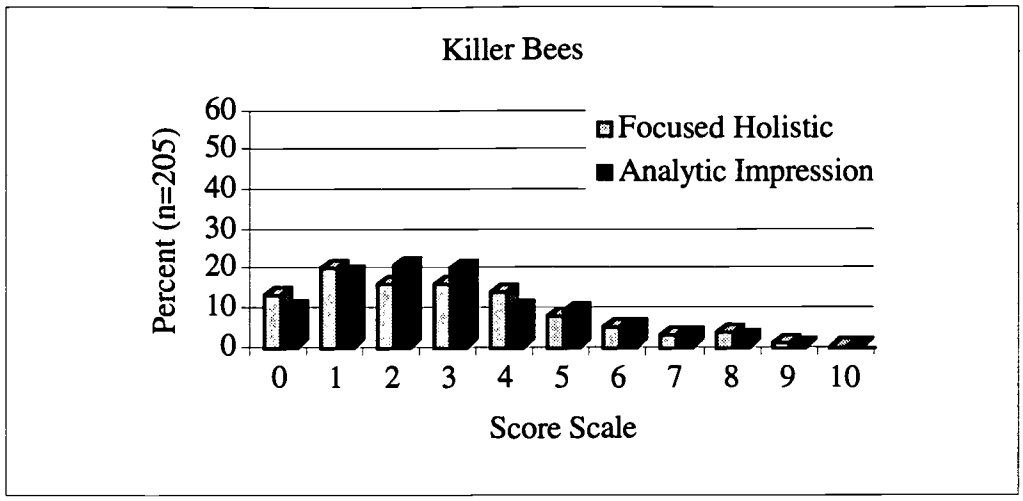
In contrast to most analytical scoring schemes when an overall score is provided, the A-I scoring method did not define how the scores on each of the separate dimensions should be aggregated in order to establish the overall score. Instead it was left to the professional judgment of each rater to decide which score point "best fit" the student's performance across each of the dimensions. For example, they were free to "average" the scores from each of the dimension or to weight one dimension more heavily than another. It is likely that, in part, this flexibility contributed to the degree of rater agreement associated with scores from the A-I method. To help understand the relationship between the scores on the separate dimensions and the overall score provided with the A-I method, scores resulting from using the A-I method to score the explanation performance tasks were further investigated by comparing the mean score on each of the four dimensions to the mean overall A-I score and calculating the intercorrelations. In addition, the degree of interrater consistency of scores within each dimension was estimated.

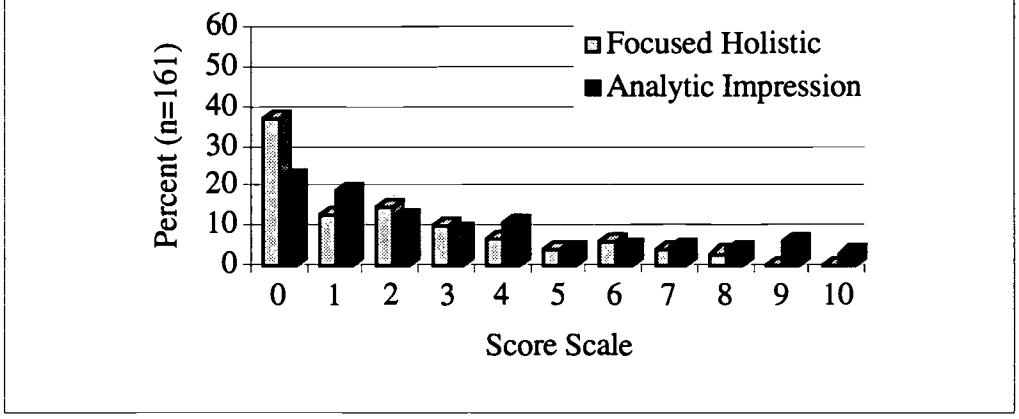
Finally, raters completed questionnaires after each scoring session to provide feedback regarding their impressions of the two scoring methods with respect to ease of use and instructional value for both teachers and students. Although 16 raters participated in at least one of the scoring sessions, feedback was obtained from only those raters that participated in both types of scoring (i.e., F-H and A-I). Of the twelve teachers who scored using both methods, questionnaires were returned by all but one of these raters.

Results

Comparison of Score Distributions

The distributions of scores resulting from the two scoring methods can be found in Figures 3 and 4, respectively, for the problem solving and explanation performance tasks. As can be seen from these figures, each of the six performance tasks was extremely difficult for the given sample of students. In particular, both of the physical science tasks (i.e., Density and Matter, and Density) were extremely difficult, resulting in J-shaped score distributions. The Earth science tasks also had J-shaped score distributions, although not as extreme as the score





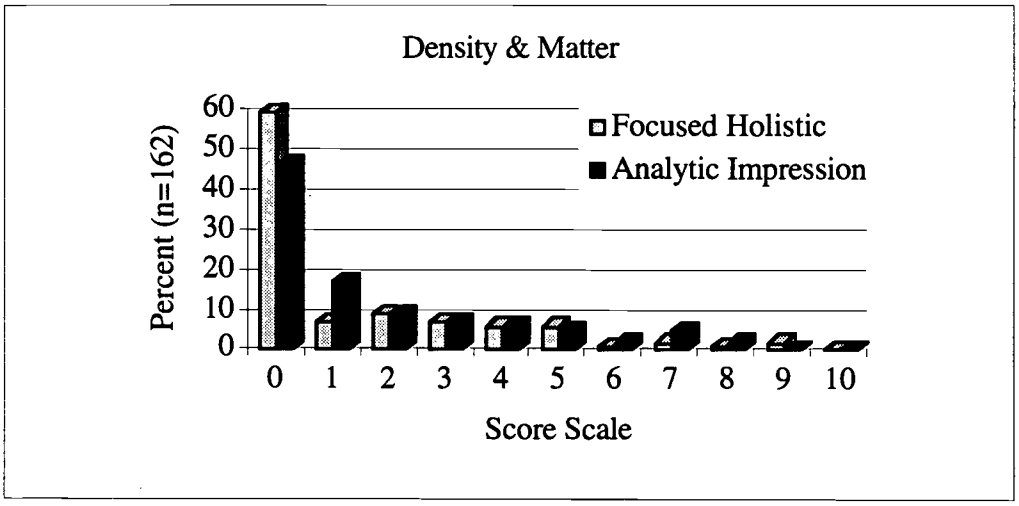
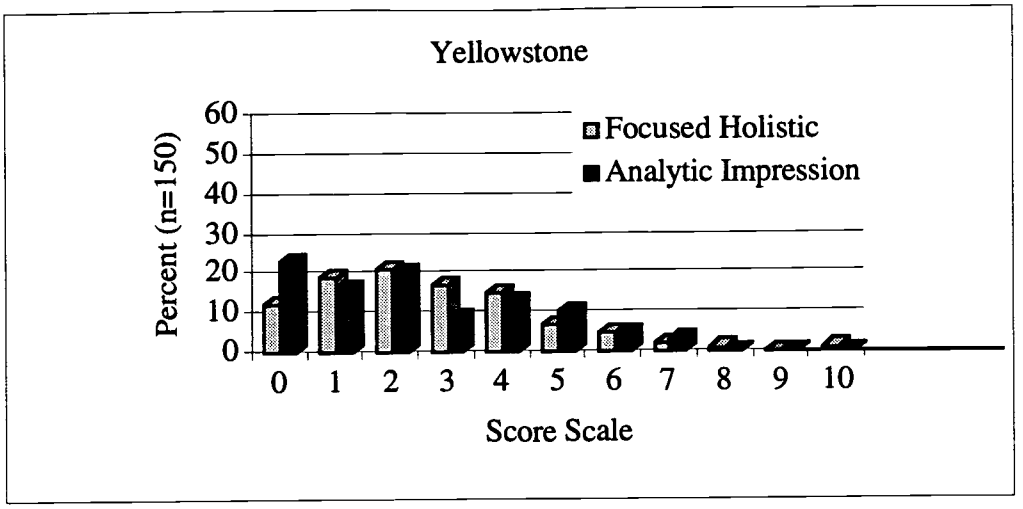
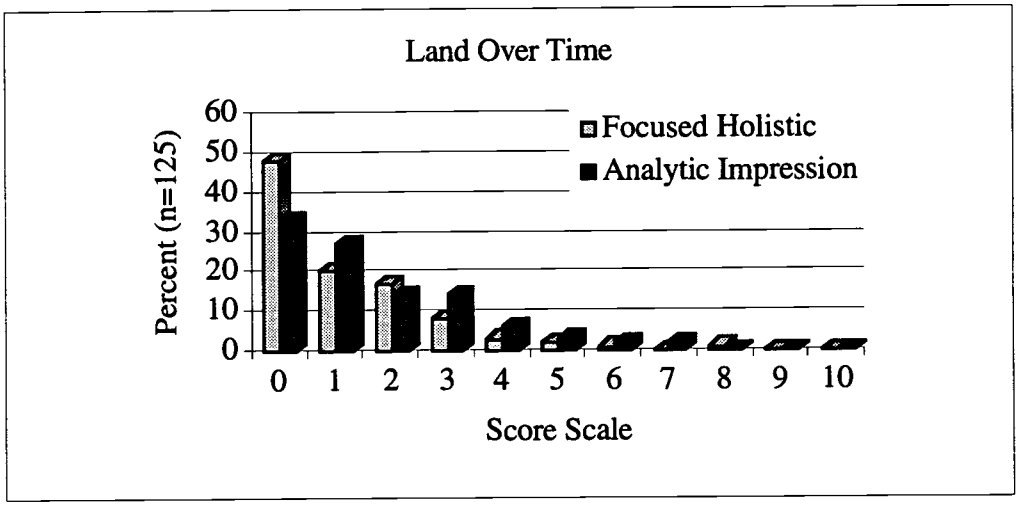


Figure 3. Score distributions of problem-solving performance tasks.





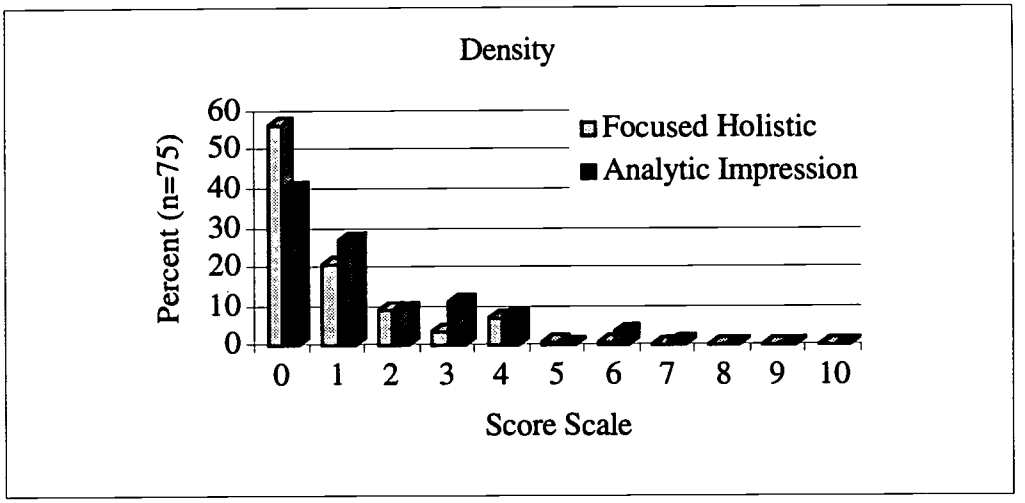


Figure 4. Score distributions of explanation performance tasks.

distributions for the physical science performance tasks. The life science tasks (i.e., Killer Bees and Yellowstone) were the least difficult of the six tasks, but even these score distributions were extremely skewed.

Although in general the F-H and A-I scoring methods produced similarly shaped score distributions for each performance task, there was one noticeable difference—the assignment of "zero." With the exception of Yellowstone, for each performance task a larger proportion of student responses received two "zero" scores from raters using the F-H method than from the raters using the A-I method. In part, this trend helps explain the difference between the means of the score distributions resulting from the two scoring methods. As can be seen in Table 3, the mean A-I scores were higher than the mean F-H scores for four of the six performance tasks. These differences were found to be statistically significant, using paired *t*-tests, at the .01 level.

Although the mean F-H scores were higher than the mean A-I scores for the remaining two tasks (Killer Bees and Yellowstone), neither of these two comparisons was statistically significant at the .05 level.

Table 3

Descriptive Statistics for Performance Tasks Using the Summed Scores

		Scoring method						
		F-H		A-I		Paired *t*-test*		
Performance task	*n*	Mean	*SD*	Mean	*SD*	*t*	(*p*)	*r*
Problem solving								
Killer Bees	205	3.0	2.26	2.7	1.93	1.84	(.067)	.72
Mountain Hike	162	2.1	2.35	3.1	3.01	-8.21	(.006)	.86
Density & Matter	162	1.5	2.27	1.7	2.29	-2.80	(.000)	.87
Explanation								
Yellowstone	150	2.7	1.95	2.4	2.03	1.94	(.055)	.73
Land Over Time	125	1.1	1.49	1.6	1.66	-5.46	(.002)	.84
Density	75	0.9	1.41	1.4	1.73	-3.24	(.000)	.66

Note. Scale = 0 to 10 points.

* 2-tailed.

In addition, the means of the F-H and A-I score distributions for the three sets of combined performance tasks (i.e., Life Science, Earth Science, and Problem Solving) were also compared using paired *t*-tests (see Table 4). For both Earth Science and Problem Solving the A-I scoring method yielded higher scores, whereas the F-H scoring method yielded higher scores for Life Science. Estimates of score reliability (i.e., Coefficient alpha) for these three combinations of performance tasks are also reported in Table 4. Although all six of these reliability estimates are moderate to low, there is a systematic difference (albeit small) favoring the F-H scoring method. Comparisons of these reliability estimates across the different combinations of tasks indicate that the two Earth Science performance tasks are more alike than are the two performance tasks comprising Life Science or the two performance tasks comprising Problem Solving.

Interrater Consistency

The values for each of the interrater consistency indices are provided in Table 5 for the six performance tasks, along with the mean agreement across tasks. Overall, the F-H scoring method yielded scores that were more consistent across raters than did the A-I scoring method. Although there is quite a bit of variability across performance tasks, raters using the F-H method typically were in exact agreement 60% of the time and agreed within one score point

Table 4

Descriptive Statistics for Combined Performance Task Scores Within Scoring Method

Combined performance task	*n*	Scoring method: F-H Mean	F-H *SD*	F-H α	A-I Mean	A-I *SD*	A-I α	Paired *t*-tests* *t*	(*p*)	*r*
Life Science[a]	96	5.4	3.52	.56	4.8	3.19	.51	2.05	(.043)	.72
Earth Science[b]	82	3.7	3.64	.70	5.2	4.27	.67	-6.86	(.000)	.88
Problem Solving[c]	98	5.1	3.67	.53	6.0	4.09	.51	-4.03	(.000)	.85

Note. Score scale = 0 to 20 points. α = Coefficient alpha.

[a] Life Science = Killer Bees + Yellowstone; [b] Earth Scienc e= Mountain Hike + Land Over Time; [c] Problem Solving = Killer Bees + Mountain Hike.

* 2-tailed.

Table 5

Interrater Consistency Coefficients Within Scoring Method

		F-H				A-I			
		% Agreement				% Agreement			
Performance task	*n*	Exact	+/-1 pt	*K*	*r*	Exact	+/-1 pt	*K*	*r*
Problem solving									
Killer Bees	205	40	87	.22	.62	38	84	.15	.45
Mountain Hike	162	60	90	.43	.76	43	83	.26	.74
Density & Matter	162	73	97	.52	.85	61	95	.38	.81
Explanation									
Yellowstone	150	50	94	.32	.68	48	84	.28	.54
Land Over Time	125	66	96	.48	.65	48	93	.04	.57
Density	75	68	95	.48	.61	53	89	.07	.64
Mean agreement		60	93	.41	.70	49	88	.20	.63

Note. Scale = 0 to 5 points. *K* = Kappa.

approximately 93% of the time. The mean Kappa for this scoring method was .41, indicating that the agreement was above that attributable to chance agreement (averaging 27% agreement over chance). In contrast, raters using the A-I scoring method typically were in exact agreement 49% of the time and agreed within one score point 88% of the time. The lower mean Kappa for this group, .20, reflects that exact agreement for these raters was much closer to chance. In fact, raters using the A-I method averaged only 14% agreement beyond chance agreement. Although less intuitive, the product-moment correlations between the sets of ratings also indicate that raters were more consistent with one another when they used the F-H method than when the A-I method was used.

When looking at the degree of interrater agreement for each task, it is evident that there is considerable variability of agreement across tasks, regardless of scoring method used. For example, Killer Bees yielded much lower exact agreement than did any other task. At first one might speculate that the lower agreement is attributable to the fact that fewer "zeros" were assigned for this set of student responses—it is much easier to agree on a "zero" than on a "one." But another plausible explanation is offered by the fact that this performance task was the first one rated by the two groups, and that perhaps the raters had not yet

gotten into their "rhythm" and were still learning how to use the scoring rubric to identify an overall score.

Another interesting comparison is the degree of interrater consistency associated with the explanation tasks scored with the A-I method. In particular, comparing the percent of exact agreement against Kappa. For example, equivalent percents of exact agreement for Yellowstone and Land Over Time (i.e., 48%) correspond to much different values for Kappa. Kappa for Yellowstone was .28, translating to 20% above chance, but Kappa for Land Over Time was only .04—only 2% above chance agreement. Even Density, where the raters agreed exactly 53% of the time, had a Kappa of only .07 (4% above chance). This phenomenon cannot be explained simply by the fact that Land Over Time and Density had more J-shaped distributions than did Yellowstone because the F-H scores for these same tasks, with more extreme J-shaped distributions, had a much higher agreement between ratings. These higher rates of agreement, however, are perhaps inflated due to the extreme number of zeros.

The level of agreement between the first ratings from the two scoring methods, provided in Table 6, indicates that the degree of agreement across methods was consistently lower than the agreement within the F-H method but higher than the agreement within the A-I method. For example, the average percent of exact agreement across methods was 52%, compared to 60% within the F-H method and 49% within the A-I method. Scores between the two methods, however, typically differed by only one score point (i.e., raters agreed with one another within one score point 89% of the time). As can be expected, when the agreement between methods was calculated using the sum of the two scores within each method the percent of exact agreement decreased significantly, with most scores differing by approximately two score points. The increased correlations between the scores from the two methods, however, are primarily attributable to the increase in score variability due to the 0- to 10-point scale.

Relationship to Test of Prior Knowledge

Based on the descriptive statistics for the prior knowledge test found in Table 7, it appears that, like the performance tasks, the test was quite difficult for the full sample of students, with the typical student correctly answering only 53% of the questions (i.e., 19 out of 36). Performance on the prior knowledge test by the subsample of students completing the separate performance tasks is also

Table 6

Interrater Consistency of Scores Resulting from the F-H and A-I Scoring Methods

		Between "first" ratings[a]				Between "summed" ratings[b]				
		% Agreement				% Agreement				
Performance task	*n*	Exact	+/-1 pt	*K*	*r*	Exact	+/-1 pt	+/-2 pt	*K*	*r*
Problem solving										
Killer Bees	205	40	82	.19	.52	34	69	91	.23	.71
Mountain Hike	162	45	80	.27	.74	37	67	84	.26	.86
Density & Matter	162	63	94	.40	.80	52	83	94	.30	.87
Explanation										
Yellowstone	150	43	89	.22	.63	26	69	92	.13	.73
Land Over Time	125	53	95	.25	.64	51	86	98	.23	.84
Density	75	65	95	.31	.61	45	75	91	.35	.66

Note. *K* = Kappa.

[a] Scale = 0 to 5 points. [b] Scale = 0 to 10 points.

Table 7

Descriptive Statistics for the Prior Knowledge Test

		Prior knowledge test[a]	
Sample of students	*n*	Mean	*SD*
Full sample	319	19.0	6.69
Problem solving			
Killer Bees	186	18.9	6.68
Mountain Hike	140	20.0	6.95
Density & Matter	144	20.5	6.89
Explanation			
Yellowstone	139	17.8	6.37
Land Over Time	111	19.7	6.41
Density	68	19.6	6.58

Note. Coefficient alpha (α) = .84. *k* = 36.

reported in Table 7. It should be noted that although the combinations of performance tasks were randomly assigned to classrooms, students responding to the Yellowstone and Killer Bee performance tasks tended to score lower on the prior knowledge test compared to the other subsamples of students. The bivariate correlations between the scores on the prior knowledge test and each set of scores for the performance tasks are provided in Table 8. These sets of correlations are quite similar for the two scoring methods with the correlations ranging from .41 to .59 and .38 to .56, respectively, for the F-H and A-I scores. Within these sets, no one scoring method consistently yielded higher correlations with scores from the prior knowledge test.

Table 8

Correlation Between Scores on Performance Tasks and Prior Knowledge Test

		Scoring method			
		F-H		A-I	
Performance tasks	*n*	*r*	$r_{\infty\infty}$	*r*	$r_{\infty\infty}$
Scores from individual tasks					
Problem solving					
Killer Bees	186	.41		.38	
Mountain Hike	140	.58		.49	
Density and Matter	144	.59		.56	
Explanation					
Yellowstone	139	.51		.41	
Land Over Time	111	.46		.49	
Density	68	.43		.51	
Scores from combined tasks					
Life Science [a]	89	.59	.86	.45	.69
Earth Science [b]	70	.56	.73	.49	.65
Problem Solving [c]	98	.66	.99	.61	.93

Note. $r_{\infty\infty}$ = disattenuated correlation coefficient,

[a] Life Science = Killer Bees + Yellowstone; [b] Earth Science = Mountain Hike + Land Over Time; [c] Problem Solving = Killer Bees + Mountain Hike.

Scores from the three combinations of performance tasks were also correlated with scores from the prior knowledge test. This set of correlations was then corrected for attenuation due to unreliability of scores from the prior knowledge test (α = .84) and unreliability of scores from the combinations of performance scores (α for each of the three combinations of performance tasks is given in Table 4). When comparing the sets of correlations associated with each scoring method, it appears that scores from the F-H method are more highly correlated with scores from the prior knowledge test than are the scores from the A-I method. This is especially true for the combination of *Life Science* tasks (.59 vs. .45). (Inspection of bivariate scatterplots and descriptive statistics indicates that the difference between the correlations for the two methods is not attributable to differences in variability, outliers, or curvilinearity.) The disattenuated correlations show a similar pattern between the two sets of correlations. One striking observation, however, is not between the two methods but rather between the three sets of performance scores. The disattenuated correlation between the *Problem Solving* scores and the prior knowledge test is .99 for the F-H method and .93 for the A-I method.

Comparison of A-I Dimension Scores to the Overall A-I Score

The mean score on each of the dimensions, for each of the three performance tasks, is plotted in Figure 5. These means were calculated by treating scores from each of the two raters as separate observations, thus maintaining the 0- to 5-point scale. As can be seen from Figure 5, the relationship of the mean scores across dimensions was similar for each of the three performance tasks. For each task, the Component Knowledge (CK) dimension yielded the highest scores, with the Communication (CM) dimension being the second highest. The Supporting Evidence (SE) and Concepts and Relationships (CR) dimensions yielded the two lowest sets of scores overall. The mean intercorrelations between these dimensions, found in Table 9, indicate that the scores on each of the separate dimensions are highly correlated with one another (.85 to .88). The scores on each of the dimensions are even more highly correlated with the overall A-I score (.91 to .93). The overall A-I score is also highly correlated (.97) with the arithmetic average of the four dimension scores, indicating that most raters used the "average" when defining "best fit."

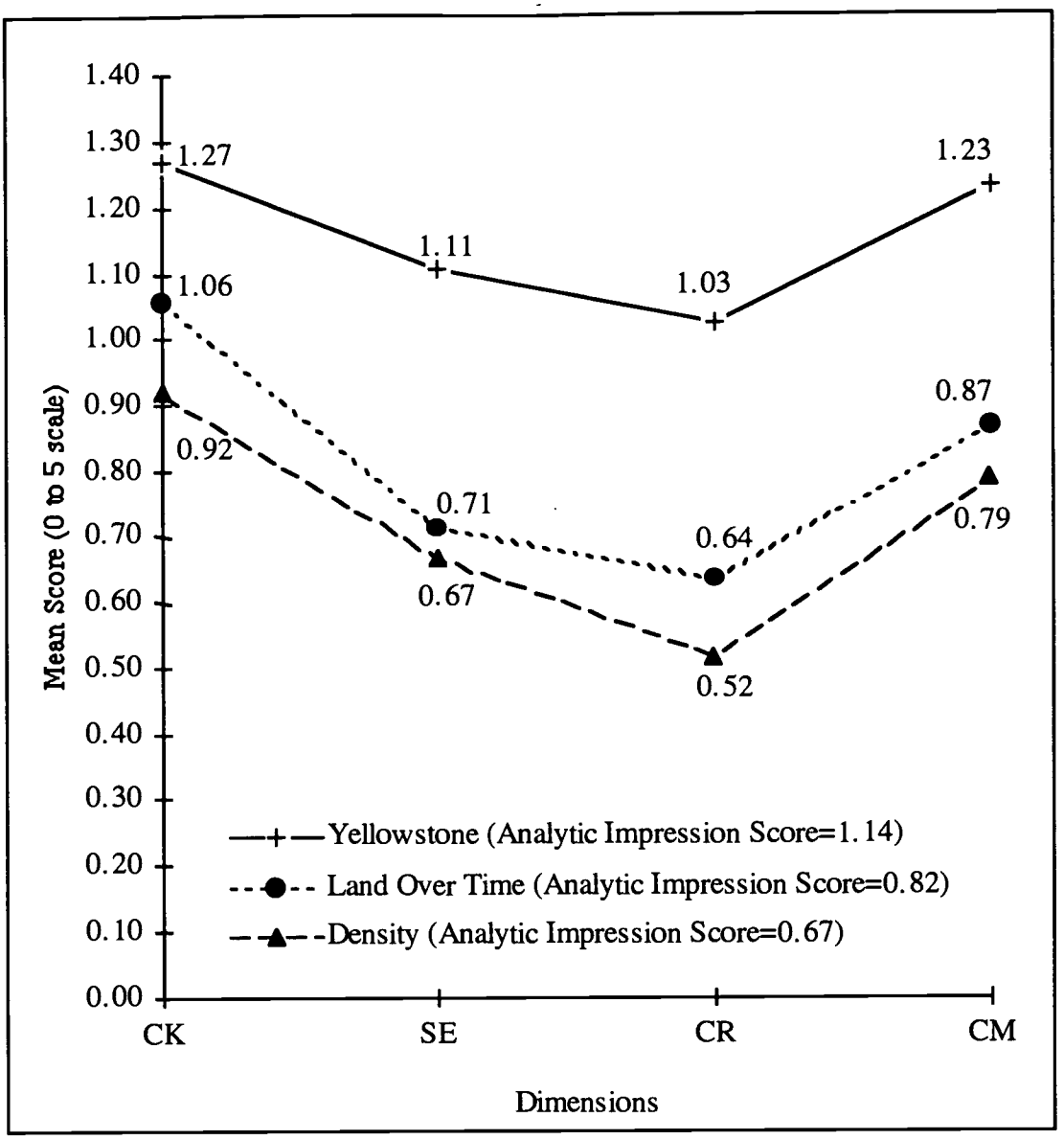


Figure 5. Mean scores for each of the dimensions using the A-I scoring method.

Interrater agreement was estimated for each of the dimension scores that were provided during the intermediate step in the A-I method in the same manner that interrater agreement was estimated for the overall scores. These estimates (i.e., percent of exact agreement, percent of agreement within one score point, Kappa, and the product-moment correlation) are provided in Table 10 for each of the three explanation performance tasks. When comparing the percent of exact agreement between raters, the CR dimension stands out as the dimension that yielded the most consistent scores. This higher degree of exact agreement, however, is probably attributable to the larger proportion of zeros assigned to

Table 9

Mean Intercorrelations Between the A-I Dimensions

Dimension	Dimensions			
	Supporting evidence	Concepts & relationships	Communication	A-I score
Component knowledge	.86	.86	.88	.91
Supporting evidence		.85	.88	.93
Concepts & relationships			.85	.92
Communication				.92
Mean score across dimensions				.97

Note. Scale = 0 to 5 points..

Table 10

Interrater Consistency Coefficients Within the A-I Dimensions

Explanation task/Dimension	*n*	% Agreement		*K*	*r*
		Exact	+/-1 pt		
Yellowstone	208				
Component knowledge		45	84	.26	.51
Supporting evidence		48	82	.28	.49
Concepts & relationships		51	80	.30	.44
Communication		48	83	.23	.49
Land Over Time	158				
Component knowledge		52	92	.29	.59
Supporting evidence		53	94	.32	.58
Concepts & relationships		58	89	.28	.54
Communication		52	94	.28	.63
Density	170				
Component knowledge		54	91	.31	.64
Supporting evidence		54	88	.19	.53
Concepts & relationships		64	92	.29	.59
Communication		47	93	.17	.53

Note. Scale = 0 to 5 points. *K* = Kappa.

student responses on this dimension. When looking across the other interrater agreement indices, there appears to be no consistent pattern that would indicate that one dimension yielded the most "consistent" scores.

In comparison with the interrater consistency associated with the overall A-I score, the consistency within dimensions (using percent of exact agreement and agreement within one score point) for *Yellowstone* and *Density* were quite similar to the consistency associated with the overall A-I score for the same task. *Land Over Time*, however, had more consistent scores within the dimensions than the percent of agreement between the overall A-I scores assigned by the raters. In addition, the Kappas within each dimension (for the *Land Over Time* & *Density* performance tasks) were higher than the Kappas associated with the overall A-I scores.

Raters' Perceptions of the Scoring Methods

Ease of use. In general, raters were divided on the method with which it was easier to assign an overall score. Raters preferring the F-H method (n = 4) stated that it was (a) faster because they only needed to make one decision rather than make separate decisions for each dimension before generating a single score; (b) more intuitive; or (c) possible to use an overall impression of quality. One rater noted that determining the overall score was particularly difficult when individual dimension scores varied by more than one point. Raters preferring the A-I method felt that it was (a) easier because it was more "clear cut"; (b) easier to separate each dimension and use that information to generate an overall score; or (c) easier to deal with responses that were not a "perfect fit."

With respect to ease of use for students in evaluating their own work, eight teachers indicated that the A-I method would be preferable. They thought this method would be less confusing for students because it is more focused. Students can concentrate on one aspect at a time rather than look at all the dimensions simultaneously. However, two teachers felt students would be more successful with the F-H rubric because it "provides guidance without the distraction of multiple decisions."

Instructional utility. Although raters did not view the two scoring methods to be substantially different with respect to their ease of use, they did perceive a difference in the potential instructional value of each method. Raters unanimously (n = 11) felt the A-I method would provide students with more

meaningful information about their own performance. They often referred to this method as precise, specific, and exact. They felt students would have a better sense of their individual strengths and weaknesses and, as a result, a clearer sense of what is expected of them. Raters (n = 9) also felt the A-I method would give teachers more insights into the effectiveness of their instructional practices. They indicated that it is easier to identify areas in which students are less prepared with the A-I method because it breaks performance down into component parts. As one rater stated, "If the students aren't addressing all the components, then maybe I'm not either." They suggested that the level of specificity obtained from the A-I scoring method would allow teachers to adjust their instruction to meet specific student needs. One rater noted, however, that the F-H method is just as useful in evaluating students' strengths and weaknesses while still providing an overall impression of performance.

Some raters also pointed out that the value of each method is very much related to the amount of time available for scoring and the intended purpose of the assessment. These raters generally felt the F-H method is most appropriate when there is a limited amount of time for scoring and the purpose of the assessment is to obtain an overall impression of student performance on a given task. On the other hand, they felt the A-I method is more appropriate when more time is available and the purpose of the assessment is to obtain diagnostic information about a student's individual strengths and weaknesses. It was also suggested that teachers who have less experience working with rubrics may do better with the A-I method. In sum, raters varied in their views about the respective ease of the two scoring methods but were in overwhelming agreement as to the instructional value of the A-I method for both teachers and students.

Overall impressions. Raters were also asked to identify (a) the approach about which they felt more confident that the scores they assigned accurately reflected the student's level of performance and (b) the scoring method they liked best. The majority of the raters (n = 7) identified the A-I method because they believed the method rubric was easier to read, allowed for compensating a student's weakness with a strength, or provided more specific information. The three raters who felt more confident in their scores using the F-H method stated that the scores were more accurate, they felt more comfortable with the method, or they felt that the method didn't force them to be heavily dependent on the

rubric. Given the above information, it was somewhat surprising that the raters were almost evenly split with respect to the scoring method they liked best. Of the seven raters preferring one method over another, four preferred the F-H and three preferred the A-I scoring method. Raters preferring the F-H method stated that it was easier to use, faster, less mechanical, or produced more consistent scores. In contrast, raters preferring the A-I method liked it because it provided more information or allowed the assignment of "extra points" for a student's strengths.

Discussion

The scores resulting from a performance task can only obtain meaning when the given performance is compared to the performance of others (i.e., norm referenced) or when the performance is compared to the criteria used in assigning the score (i.e., criterion referenced). In the context of standards-based educational reform, the importance of interpreting student performance in relationship to specified criteria has increased significantly, thus placing a stronger emphasis on understanding the relationship between the scoring criteria and the resulting scores. In this study exactly the same scoring criteria were used in both methods; only the presentation and training differed. The question then becomes, are the scores resulting from the F-H and A-I methods comparable with respect to having the same meaning in relationship to the scoring criteria?

Evidence from this study clearly suggests that the F-H and A-I scoring methods are not equally preferable for making decisions about individual students (due to low interrater agreement between the scoring methods) or even for making decisions about groups of students (due to significant differences between the mean scores for four of the six performance tasks). If the methods do not produce comparable scores, then which set of scores is better for making either individual- or group-level decisions? To answer this question, evidence from this study would suggest that scores resulting from the F-H method are preferable to those from the A-I method because of the higher degree of interrater agreement and the stronger relationship between the performance scores and the scores from the prior knowledge test. These findings, however, should be interpreted cautiously due to the extreme level of difficulty of the performance tasks.

Earlier it was mentioned that both the F-H and A-I methods could be considered to be focused holistic methods—a single score is obtained using a set of absolute criteria. In fact, the genesis of the A-I method can be traced to the point at which an early draft of the F-H scoring rubric was modified in order to better understand the distinction between the various score points. It was only by making comparisons across the dimensions that the developers of the scoring rubrics could better define each score point. Based on this experience, the question became: Would raters obtain a better conceptualization of each score point if they were trained with respect to the dimensional aspects of each of the associated criteria? Therefore, the original intent associated with the A-I method was the production of a single score, based on the rater's sound conceptualization of the scoring criteria and not the production of a score for each of the separate dimensions (although it was recognized that the rubric could be used just as easily for traditional analytical scoring).

The smaller degree of interrater agreement associated with the A-I scoring method can, in part, be attributable to the fact that the raters were not told how to combine the individual scores on each of the dimensions in order to establish an overall score. To more closely approximate the F-H scoring method, when using the A-I method it was left to the professional judgment of the raters to determine the overall score that "best fit" the student's response. Although there is evidence to support the assumption that most raters used some type of averaging system, raters combined information in different ways. For example, some raters used a student's area of strength to give the student "extra credit." This differential weighting from one rater to another (and possibly within a given rater) helps to explain the lower degree of interrater agreement associated with the A-I scoring method and explains why there were fewer "zeros" assigned with the A-I method than with the F-H method.

In addition to the technical evidence obtained in this study, feedback from the raters can be used to help make decisions regarding which method of scoring is appropriate given the context and purpose of the assessment. Although the raters were somewhat mixed with respect to which scoring method they preferred, they were overwhelmingly in favor of the A-I method when it came to obtaining useful information to improve their instruction. Also, given the current emphasis on student self-evaluation, most teachers believed that students would find the A-I method easier to use and would benefit most from

using this method. This sentiment does not come as a surprise—analytic scoring is preferred to holistic when the purpose is to obtain diagnostic information. What was somewhat surprising, however, was that even though teachers strongly preferred the A-I scoring method for obtaining information that can be used to improve instruction, the raters were equally split on the scoring method they liked best. Perhaps the increased amount of labor associated with using the A-I scoring method was a sufficient deterrent for these raters.

In conclusion, although evidence from this study indicates that the F-H scoring method is preferable to the A-I method, the extreme level of difficulty associated with the performance tasks scored in this study severely limits the generalizability of the results to other sets of performance tasks. The differences between scores from the two methods could simply be attributable to the large number of student responses receiving "zeros." It is possible that if the study were to be replicated using performance tasks that are more appropriate for the abilities of the students, the differences in the degree of interrater agreement and mean scores may dissipate. In addition, another factor limiting the findings from this study is that *generalized* rubrics (i.e., the same set of criteria is used to evaluate each performance task) were used in contrast to task-specific rubrics (i.e., the criteria are customized for each performance task). Scoring rubrics designed specifically for each performance task may function differently than generalized rubrics using the F-H and A-I scoring methods.

References

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance* assessments (pp. 19-58). Washington, DC: U.S. Department of Education.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* London: The Falmer Press.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23* , 5-12.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30* , 1-21.

APPENDIX A

SCORING RUBRICS FOR PROBLEM-SOLVING PERFORMANCE TASKS

Problem Solving
Scoring Rubric for Focused-Holistic Method

5	The response is characterized by the following criteria: - addresses all relevant constraints - all strategies and/or procedures are appropriate - implementation of strategies and/or procedures is correct - very good answer to the problem - very good understanding of the relevant skills, facts and/or concepts
4	The response is characterized by the following criteria: - addresses most relevant constraints - most strategies and/or procedures are appropriate - implementation of strategies and/or procedures is generally correct - good answer to the problem - good understanding of the relevant skills, facts and/or concepts
3	The response is characterized by the following criteria: - addresses some relevant constraints - some strategies and/or procedures are appropriate - implementation of strategies and/or procedures is partially correct - reasonable answer to the problem - fair understanding of the relevant skills, facts and/or concepts
2	The response is characterized by the following criteria: - addresses few relevant constraints - a few strategies and/or procedures are appropriate - implementation of strategies and/or procedures is generally not correct - somewhat reasonable answer to the problem - limited understanding of the relevant skills, facts and/or concepts
1	The response is characterized by the following criteria: - addresses very few relevant constraints - strategies and/or procedures are not appropriate, but there are a few appropriate elements - implementation of strategies and/or procedures is not correct - not a reasonable answer to the problem - very limited understanding of the relevant skills, facts and/or concepts
0	The response is characterized by the following criteria: - addresses no relevant constraints - strategies and/or procedures are not appropriate, or are not shown - no attempt is made at implementation - no answer is given - no understanding of the relevant skills, facts and/or concepts

Dimensions of the Problem-Solving Scoring Rubric

Understanding the problem

Understanding the problem refers to the student's ability to identify the relevant constraints specified by the problem and how these constraints relate to each other. *Constraints* refer to the conditions, variables and questions that students must adhere to while working towards a solution to the problem. One can think of constraints as the rules by which the game is played.

Planning

Planning refers to the student's ability to develop an appropriate plan of action to solve the problem. In developing this plan the student selects and organizes strategies and/or procedures that adhere to the constraints of the problem. *Strategies* refer to general plans of action which predetermine the sequence for the problem solving activity. Subgoaling, brainstorming, and looking for patterns and analogies are all examples of problem-solving strategies. *Procedures* refer to the sequence of steps or operations taken to accomplish a specific goal. If followed precisely, a procedure will invariably produce the intended result.

Implementation

Implementation refers to the student's ability to carry out the strategies and/or procedures thoroughly and accurately. For example, implementation may involve performing computations, completing a table, selecting information, generating ideas, and so on.

Answer

Answer refers to the quality of the student's answer. The answer must be consistent with the selected strategies and/or procedures and the relevant constraints.

Content understanding

Content understanding refers to the student's knowledge of relevant facts and concepts as well as the mastery of required skills. A student's content understanding is demonstrated by the ability to use relevant skills, facts and concepts correctly and consistently. Note: A student who demonstrates very good content understanding may make a minor error (e.g., a computational error, terminology error or a labeling error).

1. Problem Solving
Scoring Rubric for Analytic-Impression Method

	Understanding the Problem	Planning	Implementation	Answer	Content Understanding
5	addresses all relevant constraints	all strategies and/or procedures are appropriate	implementation of strategies and/or procedures is correct	very good answer to the problem	very good understanding of the relevant skills, facts and/or concepts
4	addresses most relevant constraints	most strategies and/or procedures are appropriate	implementation of strategies and/or procedures is generally correct	good answer to the problem	good understanding of the relevant skills, facts and/or concepts
3	addresses some relevant constraints	some strategies and/or procedures are appropriate	implementation of strategies and/or procedures is partially correct	reasonable answer to the problem	fair understanding of the relevant skills, facts and/or concepts
2	addresses few relevant constraints	a few strategies and/or procedures are appropriate	implementation of strategies and/or procedures is generally not correct	somewhat reasonable answer to the problem	limited understanding of the relevant skills, facts and/or concepts
1	addresses very few relevant constraints	strategies and/or procedures are not appropriate, but there are a few appropriate elements	implementation of strategies and/or procedures is not correct	not a reasonable answer to the problem	very limited understanding of the relevant skills, facts and/or concepts
0	addresses no relevant constraints	strategies and/or procedures are not appropriate, or are not shown	no attempt is made at implementation	no answer is given	no understanding of the relevant skills, facts and/or concepts

Dimensions of the Problem-Solving Rubric

Understanding the problem

Understanding the problem refers to the student's ability to identify the relevant constraints specified by the problem and how these constraints relate to each other. *Constraints* refer to the conditions, variables and questions that students must adhere to while working towards a solution to the problem. One can think of constraints as the rules by which the game is played.

Planning

Planning refers to the student's ability to develop an appropriate plan of action to solve the problem. In developing this plan the student selects and organizes strategies and/or procedures that adhere to the constraints of the problem. *Strategies* refer to general plans of action which predetermine the sequence for the problem solving activity. Subgoaling, brainstorming, looking for patterns and analogies are all examples of problem solving strategies. *Procedures* refer to the sequence of steps or operations taken to accomplish a specific goal. If followed precisely, a procedure will invariably produce the intended result.

Implementation

Implementation refers to the student's ability to carry out the strategies and/or procedures thoroughly and accurately. For example, implementation may involve performing computations, completing a table, selecting information, generating ideas, and so on.

Answer

Answer refers to the quality of the student's answer. The answer must be consistent with the selected strategies and/or procedures and the relevant constraints.

Content understanding

Content understanding refers to the student's knowledge of relevant facts and concepts as well as the mastery of required skills. A student's content understanding is demonstrated by the ability to use relevant skills, facts and concepts correctly and consistently. Note: A student who demonstrates very good content understanding may make a minor error (e.g., a computational error, terminology error or a labeling error).

APPENDIX B

SCORING RUBRICS FOR EXPLANATION PERFORMANCE TASKS

Explanation
Scoring Rubric for Focused-Holistic Method

5	The response is characterized by the following criteria: - very good understanding of component skills and/or facts - the explanation is well supported (the examples, facts and/or concepts are relevant, well chosen and effectively integrated) - very good understanding of key concepts and relevant relationships (no gaps or inconsistencies) - presentation is very clear and effective (notions are unambiguous and thoroughly developed)
4	The response is characterized by the following criteria: - good understanding of component skills and/or facts - the explanation is sufficiently supported (most examples, facts and/or concepts are relevant and generally integrated) - good understanding of relevant concepts and relationships (may contain a gap or inconsistency) - presentation is clear and appropriate (notions are unambiguous though they could be further developed)
3	The response is characterized by the following criteria: - fair understanding of component skills and/or facts - the explanation is somewhat supported (some examples, facts and/or concepts are relevant but loosely integrated) - fair understanding of relevant concepts and relationships (a few gaps and/or inconsistencies) - presentation is generally clear and appropriate (a few notions are vague or difficult to interpret)
2	The response is characterized by the following criteria: - limited understanding of component skills and/or facts - the explanation is weakly supported (a few relevant examples, facts and/or concepts are presented; an attempt is made to connect them, but it is unsuccessful) - limited understanding of relevant concepts and relationships (some gaps and/or inconsistencies) - presentation is somewhat clear and appropriate (some notions are vague or difficult to interpret)
1	The response is characterized by the following criteria: - very limited understanding of component skills and/or facts - the explanation is not supported (a few relevant examples, facts or concepts may be presented, but no attempt is made to connect them) - very limited understanding of relevant concepts and relationships (many gaps and/or inconsistencies) - presentation is generally unclear and inappropriate (many notions are vague or difficult to interpret)
0	The response is characterized by the following criteria: - no understanding of component skills and/or facts - the explanation is not supported (no relevant examples, facts or concepts are presented) - no evidence of understanding relevant concepts and relationships - presentation is unclear and inappropriate (communication is ineffective making the explanation very difficult to follow; or, there is not enough information to evaluate)

Dimensions Embedded in the Explanation Rubric

Component knowledge

Component knowledge refers to the discrete facts and skills underlying the concepts relevant to the task. For example, facts and skills related to understanding the concept of rectangles may include knowing that area = $l \times w$ and perimeter = $s_1 + s_2 + s_3 + s_4$, as well as having the ability to use a ruler to measure the sides of the rectangle. In evaluating component knowledge the rater should take into consideration the number of factual and/or procedural errors.

Supporting evidence

Supporting evidence refers to the examples, facts, and concepts presented to illustrate specific concepts. Supporting evidence evaluates student understanding of the relevant concepts as well as the ability to integrate ideas presented such that relationships are clearly identified.

Concepts and relationships

Concepts and relationships refers to the student's understanding of fundamental ideas (i.e., concepts) and the nature of the associations among these ideas (i.e., relationships). This dimension evaluates the thoroughness and consistency of the student's understanding of the relevant concepts and relationships as revealed through performance on the task.

Communication

Communication evaluates the ability to use relevant vocabulary and notation to effectively express ideas and describe relationships. An effectively communicated response contains language that is appropriate and unambiguous given the demands of the task and the intended audience. Furthermore, ideas and notions are developed in sufficient detail such that the intended meaning is easily interpreted.

Explanation
Scoring Rubric for Analytic-Impression Method

	Component Knowledge	Supporting Evidence	Concepts and Relationships	Communication
5	very good understanding of component skills and/or facts	the explanation is well supported - the examples, facts and/or concepts are relevant, well chosen and effectively integrated	very good understanding of key concepts and relevant relationships - no gaps or inconsistencies	presentation is very clear and effective - notions are unambiguous and thoroughly developed
4	good understanding of component skills and/or facts	the explanation is sufficiently supported - most examples, facts and/or concepts are relevant and generally integrated	good understanding of relevant concepts and relationships - may contain a gap or inconsistency	presentation is clear and appropriate - notions are unambiguous through they could be further developed
3	fair understanding of component skills and/or facts	the explanation is somewhat supported - some examples, facts and/or concepts are relevant but loosely integrated	fair understanding of relevant concepts and relationships - a few gaps and/or inconsistencies	presentation is generally clear and appropriate - a few notions are vague or difficult to interpret
2	limited understanding of component skills and/or facts	the explanation is weakly supported - a few relevant examples, facts and/or concepts are presented; an attempt is made to connect them, but it is unsuccessful	limited understanding of relevant concepts and relationships - some gaps and/or inconsistencies	presentation is somewhat clear and appropriate - some notions are vague or difficult to interpret
1	very limited understanding of component skills and/or facts	the explanation is not supported - a few relevant examples, facts or concepts may be presented, but no attempt is made to connect them	very limited understanding of relevant concepts and relationships - many gaps and/or inconsistencies	presentation is generally unclear and inappropriate - many notions are vague or difficult to interpret
0	no understanding of component skills and/or facts	the explanation is not supported - no relevant examples, facts or concepts are presented	no evidence of understanding relevant concepts and relationships	presentation is unclear and inappropriate - communication is ineffective making the explanation very difficult to follow; or, - there is not enough information to evaluate

Dimensions of the Explanation Rubric

Component knowledge

Component knowledge refers to the discrete facts and skills underlying the concepts relevant to the task. For example, facts and skills related to understanding the concept of rectangles may include knowing that area = $l \times w$ and Perimeter = $s_1 + s_2 + s_3 + s_4$, as well as having the ability to use a ruler to measure the sides of the rectangle. In evaluating component knowledge the rater should take into consideration the number of factual and/or procedural errors.

Supporting evidence

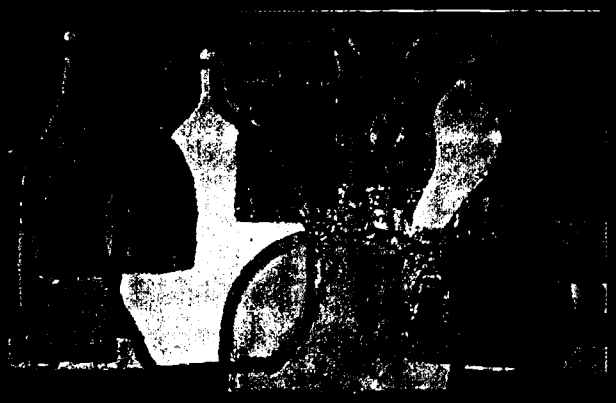
Supporting evidence refers to the examples, facts, and concepts presented to illustrate specific concepts. Supporting evidence evaluates student understanding of the relevant concepts as well as the ability to integrate ideas presented such that relationships are clearly identified.

Concepts and relationships

Concepts and relationships refers to the student's understanding of fundamental ideas (i.e., concepts) and the nature of the associations among these ideas (i.e., relationships). This dimension evaluates the thoroughness and consistency of the student's understanding of the relevant concepts and relationships as revealed through performance on the task.

Communication

Communication evaluates the ability to use relevant vocabulary and notation to effectively express ideas and describe relationships. An effectively communicated response contains language that is appropriate and unambiguous given the demands of the task and the intended audience. Furthermore, ideas and notions are developed in sufficient detail such that the intended meaning is easily interpreted.



TM029462

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").