DOCUMENT RESUME

ED 377 238 TM 022 500

AUTHOR Crehan, Kevin D.; And Others
TITLE Introducing Locally Developed Performance Measures into a School Assessment Program.
PUB DATE Apr 94
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cost Effectiveness; Curriculum Development; Educational Assessment; Elementary Education; Elementary School Students; *Elementary School Teachers; Interrater Reliability; *Reading Achievement; School Districts; *Scoring; *Test Construction; Test Use; Training
IDENTIFIERS *Locally Developed Tests; Nevada (Clark County); *Performance Based Evaluation

ABSTRACT

The Clark County (Nevada) School District introduced performance assessments into its assessment program for grades one through six as part of an effort to bring the assessment program in line with the revised curriculum. To study the effectiveness of these assessments, reading performance assessment results for grades three through five were examined, using about 100 students per grade at 11 representative schools. Assessments consisted of a short story at grades two and three and an informative text at grades four and five with questions that required students to respond in some substantive manner. Eighteen teachers from nine schools were chosen and trained as raters. Results suggest that the amount of training has an effect on rater agreement, with higher consensus after the third training session. Consensus was also greater after morning sessions than afternoon ones, suggesting an effect for fatigue. The cost of scoring these assessments was high, based on rater time, but not including the cost of training. Experience, so far, indicates that local development of a high-quality performance assessment is a formidable and expensive undertaking. It has yet to be determined if the costs are justified by the benefits. Three tables present study findings. (Contains 10 references.) (SLD)

INTRODUCING LOCALLY DEVELOPED PERFORMANCE MEASURES INTO A SCHOOL ASSESSMENT PROGRAM

Kevin D. Crehan

University of Nevada, Las Vegas

Rhoton Hudson

Judith S. Costa

Clark County School District

PROBLEM

During 1992 and 1993 the Clark County (Nevada) School District's (CCSD) Department of Testing and Evaluation conducted a thorough revision of its curriculum-based assessment program in grades one through six. The revision was necessary to bring the assessment program in line with revised curricula and to supplement the revised multiple-choice tests with performance assessments. The introduction of performance assessments was in response to the growing interest in more "authentic" assessment associated with national concerns for educational reform. Following intensive preparation and field testing activities, the new tests

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 1994.

were administered for the first time in the spring of 1993. Also for the first time, concerns related to scoring the written student responses were presented. Among the questions were:

What level of rater reliability is satisfactory?

How much training is necessary to obtain satisfactory reliability?

Should raters be screened for agreement?

How many raters should score each response?

How much is all this going to cost?

Since there are over 70 thousand students in grades one through six in the CCSD, the cost of administering and scoring an individual performance assessment is an important consideration.

REVIEW

An in-depth analysis of the questions involved in the implementation of performance assessments is presented by Linn, Baker, and Dunbar (1991). They counsel that the issue of evaluating the quality of alternative assessments has not been sufficiently considered and suggest eight criteria for judging the adequacy of performance-based assessments: 1) consequences, 2) fairness, 3) transfer and generalizability, 4) cognitive complexity, 5) content quality, 6) content coverage, 7) meaningfulness, and 8) cost and efficiency.

While all criteria are important and addressed in some manner by CCSD, the present focus is

on criteria eight and three, i.e., cost and quality of scoring. In a recent review, Linn (1993) reports satisfactory generalizability across raters has been observed in a number of contexts given explicit scoring rubrics with intensive reinforced training. Additionally, the California Assessment Program has established an inter-rater reliability of .90 for their writing assessment by using procedures which include providing sample anchor papers for each rater and recirculating previously scored papers to check on stability (U.S. Congress, Office of Technology Assessment, 1992). Shavelson, Baxter, and Pine (1992) observed the reliability and validity of performance assessments in the 5th and 6th grade science curriculum. They asked the question: How large a sample of observers is needed to produce reliable measurement? Their results found inter-rater reliability to be consistently high in evaluating student performance on complex tasks, high enough to conclude that a single rater provides a reliable score.

While the reports of Linn (1993) and Shavelson et al. (1991) are promising, earlier writers are less encouraging. In reviewing the pros and cons of essay examinations, Coffman (1971) reports a lack of conformity in scoring among different raters. Coffman and Kurfman (1968) found two raters differing by 142 points on a set of 60 papers, which suggests that, if a specific score is needed to pass an examination, then the severity of the person scoring the paper will determine whether it passes or fails. Coffman also found that raters can vary in how they distribute grades across the score scale and in the value they place on different papers as well as in how strictly they score. In his review he observed inter-rater reliability coefficients ranging from .35 to .98, depending on the context, content, or number of raters

scoring. Godschalk et al. (1966) found that essay examinations read toward the end of a several day scoring session tend to receive lower scores than those read earlier in the grading session. Training included rating sample papers and comparing scores with scores given by other raters. For a large field test, the inter-rater reliability was only .672 for three readers.

Moss (1992) and Linn (1993) observed that there is a problem concerning comparability of scores assigned by different raters. This source of error is attributed to the necessity of reliance on professional judgment in scoring performance assessments. However, Linn (1993) notes that, with careful training of raters on well designed rubrics, the error variance due to raters is less than that due to task specificity. In reviewing data from Baker (1992) and Lane, Stone, Ankenmann, & Liu (1992), Linn shows substantially greater increases in score generalizability from increasing the number of tasks than from increasing the number of raters.

RESEARCH QUESTIONS

It can be concluded from the review that, although measurement error due to raters is a concern, acceptable inter-rater reliability is attainable given appropriate reinforced training and carefully prepared scoring rubrics. Unfortunately, the practical question of how much training is necessary to ensure satisfactory reliability for a specific context and specific situation could not be adequately answered by the research of others. Therefore, the present

study was conducted to estimate the effect of the amount of training on inter-rater agreement and performance score generalizabilty as well to estimate the costs of training and scoring in the local setting.

METHODS

The reading performance assessment results for grades two through five from eleven representative CCSD elementary schools with approximately 100 students per grade per school were used in the study. Eighteen teachers from nine different elementary schools were chosen as raters for the study. None of the teachers had any previous training or experience in scoring these performance assessments. The teachers were divided into four groups, with five teachers each in the second and third grade scoring groups and four teachers each in the fourth and fifth grade scoring groups. On the first day of the study, after the teachers were assigned to a grade level scoring group, they were given a stack of assessments at their grade level selected randomly from each of the eleven representative elementary schools. The performance assessment consisted of a short story at grades two and three and an informative text at grades four and five with questions that required the students to respond to the passages in some substantive manner. They were also given a scoring rubric which showed how to determine the score to give each assessment. Lastly, they were given a set of anchor papers for reference. The anchor papers were actual assessments which showed examples of papers scored at each of the possible rubric scores. In grades two and three, the scores could range from 0, indicating the assessment answer

was missing or inappropriate, to 3, which meant that the student had accomplished all the requirements at a mastery or competency level. In the fourth and fifth grades, the scores could range from 0 to 4, with 4 being mastery level.

The raters were instructed to grade their stack of tests to the best of their ability with no formal training on how to use the rubric and anchor papers. After approximately an hour and a half of scoring student papers with virtually no training, all scorers reconvened for the first session of formal training. During the first hour of their first training session, the raters from all four groups were together for general training. The two trainers gave all raters the same rubric and anchor papers. Differences among the rubric scores were explained, supported by the anchor papers. After an hour, the raters were divided into two groups, with second and third grade teachers together, and fourth and fifth grade teachers together. For the next 45 minutes, training continued with reference only to anchor papers at the specific grade levels the teachers would be scoring. After this hour and 45 minutes of training, the raters were given another set of assessments to score. When they finished this set, they adjourned for the day.

At the beginning of the second day, each group of raters was given a 45 minute refresher orientation before the third scoring session. During this training, the raters went over actual assessments that they had scored the day before. They discussed some papers on which they had all reached consensus, other papers where the raters had been one score point different, and some papers where there was a high degree of divergence. After the 45 minute training

sessions raters again split into groups and spent approximately an hour scoring. After a break for lunch, the raters were given additional training of one hour and 20 minutes on papers they had scored during session three, followed by two additional hours of scoring.

The third day started with one hour of training at each grade level, using tests from the scoring session at the end of day two. During this training, all raters came to consensus on what the score should be for each paper and all said they felt comfortable that they could score similar papers the same way. Raters were reminded that the length of a student's response is not necessarily related to the quality of the response. They were also reminded that the writing mechanics in a paper should not be considered in determining the reading score. Raters scored for one and a half hours, and then reconvened for a quick final 15 minute period to compile and compare their scores from session five. Discussion focused on papers where score divergence was present. The second-, third-, and fifth-grade raters then participated in a final scoring session of one hour 15 minutes.

RESULTS

Summary results of training, scoring, and percent agreement among raters are presented in Tables 1 and 2.

Insert Tables 1 and 2 about here

Scores recorded during the hour and a half pretraining scoring session showed that all raters

at each grade level either reached consensus or scored within one point 76 percent of the time at second grade, 53 percent at third grade, 36 percent at fourth grade, and 84 percent at fifth grade. (It should be noted that while the fifth-grade raters scored nineteen papers, only six papers were read by all four raters due to a procedural error, while second-, third-, and fourth-grade raters scored an average of 20 of the same papers during the pre-training session.) It took the raters an average of 4.25 minutes to read and score each assessment. It took the fourth- and fifth-grade group slightly longer than the second- and third-grade group because the upper grades' passages were longer, and the students' responses were more involved.

Following the first formal training session of approximately one and three-quarters hours and the subsequent one-hour-long scoring session, agreement among scorers showed an increase in all but fifth grade. The time required to grade each test also dropped at each grade level to an average of slightly less than two minutes per test.

The third scoring session (which followed two formal training sessions) showed an increase in agreement in all grades except the second, where inter-rater agreement dropped slightly. The length of time required to grade each test was again a little less than previous sessions with the average at a rate of about one test every 1.5 minutes.

Results of the fourth scoring session (after three training periods) actually revealed an overall decrease in agreement among raters. Rating which were the same or within one point dropped substantially in grade three and slightly in grades four and five. The amount of time

required to score rose to nearly four minutes per test. This decline may be attributable to the fact that it was late in the day and that raters scored for two straight hours, which was the longest scoring session without a break.

The fifth scoring session (the last in which all raters at all grade levels participated) resulted in a substantial increase in agreement compared to the previous session. Raters agreed within one point in 83 percent of the assessments at grade three, 85 percent at grade five, 90 percent at grade two, and 96 percent at grade four. The amount of time spent on scoring each test had dropped back to an average of about two minutes each.

Results of the final scoring session, which followed a brief 15 minute reinforcement training period, showed large declines in rater agreement in second and third grades.

As an additional indicator of scoring consistency a sample of seven to nine assessments from each grade level, that had been previously scored by three "expert" raters (who agreed on the score), were inserted in assessments to be scored, with one or two of the assessments included in those to be scored after each training session at each grade level. With the exception of one assessment at third grade, all raters scored within one point of the original score on all assessments regardless of the session in which it was scored.

In addition to percent agreement among scorers reported in Tables 1 and 2, G-study coefficients of generalizability were determined for each session and grade level. These

results ape the percent agreement results

Insert Table 3 about here

fairly closely. Unfortunately, the magnitude of score generalizability is modest in most instances. However, given that there was only one task rated, the G-study coefficients compare quite favorably with those in other studies (e.g., Baker, 1992).

DISCUSSION

Results indicate that the amount of training seems to have an effect on rater agreement. Overall, raters showed higher consensus after each morning training-scoring session with declines in agreement observed in the afternoon sessions. The greatest gain was for the third scoring session when raters had received a total of about two and one half hours of training. The size of the time of day effect was somewhat surprising and can only be explained by fatigue.

Finally, the cost of scoring this type of performance assessments is high, as has been noted in the literature. A conservative estimate, based on a rater scoring an average of 40 assessments an hour at $20 per hour, is $.50 per test. This figure does not include the additional expense involved in training, developing and administering the assessments. Additionally, the present scoring task was to score only one aspect of the performance

assessment, response to reading. The complete rubric also includes a score for language mechanics. Therefore, in practice, the cost per assessment would be higher.

In summary, based on the experiences and learnings thus far, it is concluded that local development of a high quality performance assessment is a formidable (and expensive) undertaking. It is yet to be determined if the costs are justified by the benefits.

REFERENCES

Baker, E.L. (1992). The role of domain specification in improving the technical quality of performance assessment (Tech. Rep.). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Coffman, W.E. & Kurfman, D.A. (1968). A comparison of two methods of reading essay examinations. American Educational Research Journal,5,99-107.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.) Educational Measurement (2nd Ed.). Washington, D.C.: American Council on Education.

Godshalk, F.I., Swineford, F., & Coffman, W.E. (1966). The Measurement of Writing Ability. New York: College Entrance Examination Board.

Lane, S., Stone, C.A., Ankenmann, R.D., & Liu, M. (1992). Empirical evidence for the reliability and validity of performance assessments. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher,20,15-21.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis,15,1-16.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research,62,229-258.

Shavelson, R. J., Baxter, G. P., & Pine J. (1992). Performance assessments: Political rhetoric and measurement reality. Educational Researcher,21,22-27.

U.S. Congress, Office of Technology Assessment. (1992) Testing in American Schools: Asking the Right Questions (OTA-SET-519), Washington, D.C.: U.S. Government Printing Office.

TABLE 1

SUMMARY OF TRAINING, SCORING AND PERCENT AGREEMENT

	SCORING SESSION	PRIOR TRAIN	TIME SCORE	NUMBER SCORED	SAME SCORE	1POINT DIFFER	2 POINT DIFFER
GD 2	1	0	85	29	24%	52%	24%
	2	105	60	30	40%	45%	15%
	3	45	60	41	39%	33%	28%
	4	80	120	38	33%	42%	25%
	5	60	90	51	35%	55%	9%
	6	15	75	58	41%	38%	20%
GD 3	1	0	85	21	5%	48%	38%
	2	105	60	41	27%	44%	29%
	3	45	60	58	31%	58%	10%
	4	80	120	33	6%	66%	18%
	5	60	90	52	21%	62%	16%
	6	15	75	60	17%	45%	38%
GD 4	1	0	85	11	0%	36%	64%
	2	105	60	33	9%	39%	52%
	3	45	60	33	24%	63%	12%
	4	80	120	20	45%	40%	15%
	5	60	90	39	67%	29%	5%
GD 5	1	0	85	19	0%	84%	17%
	2	105	60	32	21%	58%	21%
	3	45	60	41	19%	69%	11%
	4	80	120	32	6%	81%	12%
	5	60	90	35	31%	54%	16%
	6	15	75	46	17%	76%	7%

TABLE 2

PERCENTAGE OF RATINGS WITHIN ONE POINT

	SESSION					
	1	2	3	4	5	6
Grade 2	76	86	71	75	90	79
Grade 3	53	71	89	72	83	62
Grade 4	36	48	87	85	96	--
Grade 5	84	79	88	87	85	93
MEAN	62	71	84	80	88	78

TABLE 3

G-STUDY COEFFICIENTS

	SESSION					
	1	2	3	4	5	6
Grade 2	.58	.81	.53	.79	.79	.58
Grade 3	.42	.80	.77	.52	.68	.56
Grade 4	.21	.61	.64	.81	.83	--
Grade 5	.70	.60	.68	.55	.66	.75
MEAN	.48	.70	.66	.67	.74	.63