DOCUMENT RESUME

ED 358 146 TM 019 935

AUTHOR	McColskey, Wendy; And Others
TITLE	Matching Assessment to the Curriculum in a Systemic Science Reform Project.
PUB DATE	Apr 93
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Curriculum Development; Curriculum Evaluation; Educational Assessment; *Educational Change; Educational Objectives; Elementary Education; *Elementary School Science; Elementary School Students; Evaluation Methods; Grade 6; Intermediate Grades; Middle Schools; *Science Education; Science Instruction; Scores; *Student Evaluation; Systems Development; Test Construction; *Test Use; Test Validity
IDENTIFIERS	*Curriculum Based Assessment; Middle School Students; National Science Teachers Association; North Carolina; *Reform Efforts

ABSTRACT

The North Carolina Project for Reform in Science Education, which is part of the National Science Teachers Association's Scope, Sequence, and Coordination Project, and the importance of student test scores as an evaluation tool were evaluated. Seven North Carolina schools participated in the project's first year (1991-92), with 21 teachers, and about 1,600 sixth graders. The reform program called for science for 6 years for all students, fewer topics covered in mode depth, and a careful and coordinated curriculum with a hands-on approach and assessment examining understanding. Because the goals of the reformed curriculum differed from the curriculum on which the state-mandated science test was based, a method was developed to adjust test scores to allow for the difference and allow the test to be instructionally valid. The adjustment process protected project schools from uninformed decisions about project worth. Students in project schools performed slightly better than did students in control schools on subject matter taught more extensively in the project schools. Efforts to develop alternative assessments for the project schools are briefly described. Results highlight difficulties arising in using student testing as a program effectiveness summative criterion. Experiences show some legitimate formative functions that tests can serve. Appendix A contains a matrix of sixth grade goals by content area. Appendix B gives sample performance assessment item sets. (SLD)

Matching Assessment to the Curriculum in a Systemic Science Reform Project

by

Wendy McColskey
Patrick Harman
SERVE

Helen Weaver
East Carolina University

Jim Fortune
Virginia Tech

Paper Presented at the American Educational Research Association Conference, April 14, 1993, Atlanta, Georgia

Matching Assessment to the Curriculum in a Systemic Science Reform Project

Introduction

"Is this science education reform effective?" is a question many project funders, managers, and evaluators are struggling with as science education reform has come to the forefront of many professional, state, and local agendas. Student testing is the traditional way to answer the question of program effectiveness. That is, traditionally, the main purpose of student testing in a science curriculum reform evaluation is to determine if students have learned more than they would have otherwise without this experience. Accepted practice is to dismiss negative findings and focus on significant differences.

Perhaps we need to re-examine the role of student test data in an evaluation effort. Stake (1985), on the basis of experience with many evaluations, argues that gain in student performance is a weak approximator of the quality of the program. He stated:

> "Student performance data are important information to those responsible for the development of innovative programs, but we could find no justification for treating such data as program-effectiveness criterion data in most evaluative studies (even though an RFP might specifically define them so)."

Our experience with evaluating the National Science Teachers Association's Scope, Sequence, and Coordination project in North Carolina leads us to a similar conclusion. Student performance data are important and serve several important functions in a reform project, but perhaps, test scores have been oversold as a measure of the worth of a program. Although summative uses of student performance data may be less than beneficial, formative uses can be very powerful if these uses model reflective thinking about goals and encourage discussions among practitioners involved.

The Program

The North Carolina Project for Reform in Science Education (NCPRSE) is part of the of the National Science Teachers Association (NSTA) Scope, Sequence, and Coordination initiative, located in five states and Puerto Rico, which consists of programs aimed at reforming science education. Some of the goals of the NCPRSE curriculum are to improve student attitudes toward science, to increase student interest in science careers especially for females and minorities, to increase student critical thinking skills, and to improve student performance in science. The strategies to be employed are those recommended by the National Science Teachers Association:

1) all students study science every year for six years

2) fewer topics are studied more in depth

3) concepts from major disciplines are studied every year in a carefully sequenced and coordinated fashion (rather than in a layer-caked fashion)

4) hands-on experiences come first before abstract concepts

5) the curriculum assesses depth of understanding not just facts or information.

Seven North Carolina middle schools participated in the project in 1991-92. There were twenty-one teachers and approximately 1600 sixth grade students who participated in the project's first year.

State-mandated tests as evidence of project success

Context: In a state environment in which policy-makers and the public have come to expect to see published reports of school science test scores, these tests become measures of project success that must be dealt with. In this case, the state-mandated sixth grade science test was not instructionally valid for the NCPRSE curriculum. That is, the reform curriculum goals were quite different from those of the state curriculum upon which the test was based. Consequently, the possibility existed that project school scores on the state-mandated, end-of-year science test would decrease from the previous year due to this mismatch and that the project might be stopped by some local administrators because of this decrease. The state department, because of its support for science reform projects, agreed to work with the project to devise a method to adjust test scores so that project schools would not be penalized for curriculum objectives and test content not covered. Because test results were available and looked to as a measure of success, the project evaluation included it as an outcome measure but used the adjusted scores to make the test more instructionally valid.

Method: The adjustment of scores consisted of the following: 1) developing a rating scale such that project teachers could rate the items on the test in terms of whether or not they had covered the content or skill with their students; 2) showing the sixth grade science test (60 multiple choice items) to the teachers after the test was administered and having them rate each item; 3) obtaining student test data from the state department; 4) multiplying a "1.1" (correct response) or "0.1" (incorrect response) for each item times the particular student's teacher rating of the item; 5) recalculating the total score and rescaling this new total score back to the original scale; 6) comparing project schools to control schools on adjusted and unadjusted scores; 7) providing project school adjusted score averages to the state for consideration in computing district averages in annual state reporting.

Findings:

1) As expected the project teachers covered fewer items with their students than teachers in control schools.

Table 1: Comparison of mean number of items rated as having been covered, by teacher type, subscale, and total

	Subscale (12 items in each)					
	Life Science	Physical Science	Earth Science	Nature of Science	Process	Total (60)
Project	3.86	5.68	4.52	4.81	6.33	25.19
Control	8.58	9.21	6.74	9.11	9.05	42.68

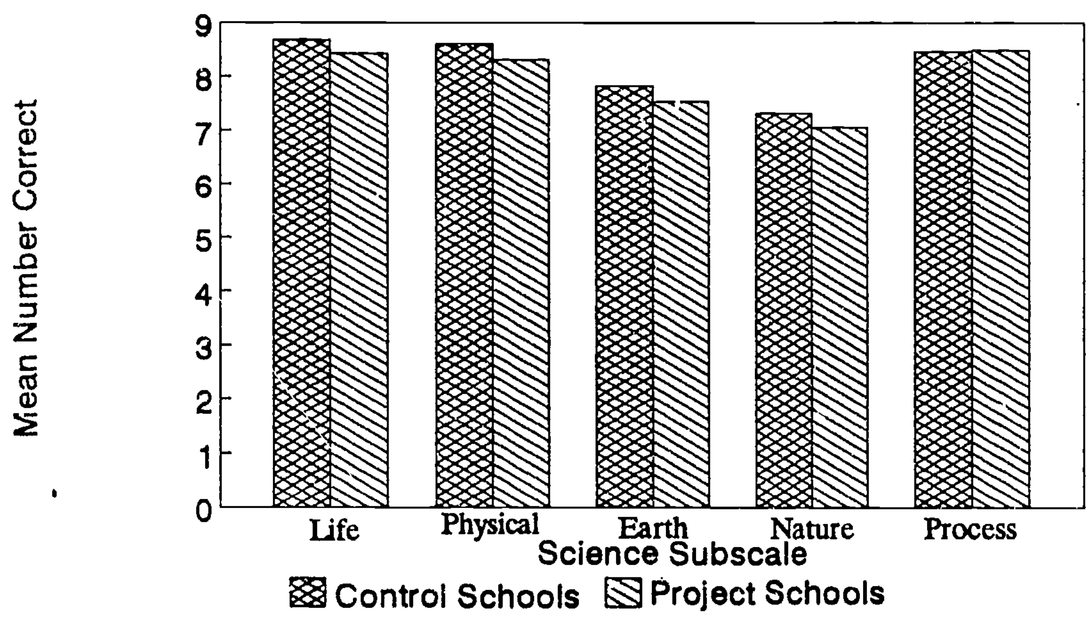
2) The comparison of predicted science scores (based on scores obtained in language arts, math, and social studies) to actual scores shows that actual scores for the project group were slightly depressed (39.84) relative to their predicted scores (40.94). The control schools' scores (actual) were statistically higher than project school scores although the mean difference was approximately one point. After the adjustment, the project schools scored slightly, but not significantly higher than the control schools. Thus, the adjustment process "protected" project schools from uninformed decisions about project worth.

Table 2: Results of two ANCOVAs Comparing Science Test Raw Score Means for Unadjusted and Adjusted Scores with CAT as Covariate, by School Type (60 item test)

Data Type	Program	Comparison	F ratio	p level
Predicted	40.94	40.88	N/A	N/A
Unadjusted	39.84	40.88	9.13	.0047
Adjusted	43.01	42.04	0.88	.3540

3) Our analyses, summarized in Figure 1 below, showed that students in control schools did slightly better than students in the project schools on subject matter content taught more extensively in the control schools. We did not find any difference in performance on a subscale of process items, which are more closely aligned with the curriculum goals of the project schools.

Figure 1.
Mean Number Correct
by type and subscale (12 items each)



Discussion:

The results detailed above aid all those involved in better understanding the relationship between the state test and the program and provide some information on the degree to which teachers felt free to ignore state curriculum goals, but do not provide evidence of project success. In a review of the effects of NSF-funded curriculum projects initiated in the 60s, Welch (1979) found that of the 115 criterion measures used in comparing project to control schools, 58 favored the "new" curriculum. It is unlikely that the remainder of these programs were bad programs. It is more likely, as in this evaluation, that the performance sample was too small or other important factors not related to the curriculum were too large to demonstrate conclusive effects from the curriculum reform.

In a "high stakes" state testing environment, teachers take seriously their responsibility to teach to the state curriculum goals and test. A reform which espouses different goals puts teachers and administrators in a difficult position. An adjustment process educates those in the field that no test is valuable in and of itself, without consideration of its instructional validity. It is only in its match to the curriculum goals that student performance information from tests becomes valuable. When teachers are asked to rate test items, they see more clearly the inadequacies of any one event, multiple choice test as a summary or judgment of their efforts. Thus, they are more likely to be empowered to teach to the new goals.

In conclusion, we think examining state-mandated test score data although not powerful as evidence of project success can:

1) encourage schools to participate in a project with goals that differ from state curriculum goals;

2) protect against decisions to drop the program because test scores went down;

3) empower teachers to examine the instructional validity of state tests and consider their meaningfulness for their purpose;

4) provide information about how willing teachers have been to let go of teaching to the test.

Developing alternative assessments as evidence of student success

Context: The state-mandated science test represents a traditional, on-demand, small item sample, multiple-choice test. Thus, there was interest in the project in trying to develop "alternative assessments" that could describe student achievement relative to curriculum goals. However, several problems arise in trying to develop and use alternative assessments for summative purposes in an evaluation effort. These are:

1) It may take several years for curriculum goals, instruction, and assessment methods in the classroom to coalesce in a way that leads to some consensus on what is expected of students.

2) Teachers, in the initial years of a reform, are so busy with the day-to-day, practical issues involved in changing instruction that there is little time left for involvement in the development of assessments for program evaluation.

3) The evaluation resources needed for the development, administration, and scoring of "alternative assessments" are beyond those available in most evaluation budgets.

4) Even if carried out, interpreting alternative assessment data in a way that is convincing to the public and funders creates new challenges.

Given these problems, we decided to develop alternative assessments, but to use them in formative ways.

Method: We found that the process of walking through this developmental process in conjunction with a developmental curriculum effort can lead to enhanced thinking about project goals. The "alternative assessment" development process included the following:

1) The curriculum developer provided a rough matrix of program goals by content areas (Appendix A).

2) Assessments were developed by sixth grade science teachers not involved in the project.

3) Early forms of the assessments were shared with project teachers and piloted in several project teachers' classes.

4) Assessments were administered to a sample of students at each project school with project teachers present. Project teachers then reviewed the work and discussed the quality of the samples produced. (Appendix B contains two of the item sets used.)

5) Project teachers were asked to predict the percent of students who would perform at three different levels (would fully achieve, partially achieve, and not achieve) on the objectives assessed.

6) The project evaluator roughly sorted the student responses into three groups (incorrect, partially correct, and correct) and discussed the results with teachers (See Table 3).

7) Teachers were interviewed to find out their perception of the importance of various curriculum goals and how or if they formally assessed student progress on each. The purpose of these interviews was to discover the match between stated curriculum goals used in developing "alternative assessments" and goals being pursued by teachers in the classroom.

Limitations in terms of resources necessitated using convenience samples of students and one rater (the evaluator) of student responses.

Although not good practices from a research perspective, the test information itself was less important than the developmental process itself. The benefits may be greater if assessment is taken on as an educative, participative process of describing student achievement, rather than a summative process of proving project worth.

Sample finding: The following chart summarizes some of the results on open-ended questions:

Table 3: Results on items relating to the NCPRSE goal that students will be able to explain relationships and basic concepts related to the apparent motion of the sun.

	Fully Achieved	Partially Achieved	Not Achieved
Avg. Teacher Expectation	**63%**	**26%**	**11%**
Open-Ended "Explanation" Questions:			
1. Explain Tilting	58%	14%	28%
2. Explain Sun Moving	53%	15%	32%
3. Explain Shadows	32%	29%	39%
4. Explain 9:00-4:00	50%	16%	34%

Median Sample Size: 111 students across seven schools

Discussion:

Although teachers found the information provided in the above chart very interesting, it is unlikely to be the kind of information that would convince others that the reform was effective. From these assessments, several teachers realized the difficulty many students have in explaining or articulating their understanding of science concepts. That is, students may be able to recognize right answers on a multiple-choice or fill-in-the-blank tests, but coherent, accurate explanations are only produced by about 50% of the students assessed (a smaller percentage than the teachers expected).

Our thinking upon reflecting on this process is that it is not so much "test information" by itself that is critical, but it is the conversations and learning that arise out of the development process that are so important. Thus, rather than having the evaluator own the assessment data, it may be important to encourage site-based evaluation.

Conclusions

Our experience with evaluating the North Carolina Project for Reform in Science Education has demonstrated the difficulties arising in using student testing as a program effectiveness summative criterion. Specifically, previously developed tests (in this case, state mandated) are not good matches to curriculum reform goals and thus, must be adjusted to be more instructionally valid. In addition, major curriculum effects on student learning measures are seldom found due to the many other important factors that affect this outcome. Our efforts to develop alternative assessments designed to more accurately reflect the goals of the reform project demonstrated that development of alternative assessments are time-consuming and resource intensive to develop and thus, not practical for most evaluations.

Our experience also highlighted several legitimate formative functions which can be served by evaluator's testing efforts. The use of a state-mandated test provides accountability by adjusting test scores and allowing a reform to get off the ground. That is, project schools may not participate unless they feel protected from state accountability systems. Using an adjustment process empowers teachers to consider the validity of tests by having them rate items. Our use of alternative assessments pushed the discussion and articulation of curriculum goals as a precursor to the development of "alternative assessments" and involved teachers in an outcomes assessment process, at least as observers, which can transfer to classroom and school site program evaluation practices.

REFERENCES

Stake, R.E. (1985). A Personal Interpretation. Educational Evaluation and Policy Analysis, 7(3), 243-244.

Welch, W. (1979). Twenty years of science curriculum development: A look back. In Review of Research in Education, Berliner, D. (ed.). Washington, DC: AERA.

APPENDIX A

MATRIX OF SIXTH GRADE GOALS BY CONTENT AREAS

NCPRSE ASSESSMENT MATRIX

Goals	Content											
	Apparent Motion	Shadows	Direction	Soil	Weather	Location of Water	Water Phases	Water Cycle	Biomes	Organisms	Sun	Light
To use simple instruments					Grp/ Manip #5							
To follow directions	Ind/P&P #2	Ind/P&P #3	Ind/P&P #1		Grp/Manip #5							Self/Ma-nip #4
To ask questions												Self/Ma-nip #4
To collect or record data		Ind/P&P #3			Grp/ Manip #5							
To maintain a journal												
To comm-unicate observations	Ind/P&P #2	Ind/P&P #3	Ind/P&P #1		Grp/ Manip #5							Self/ Manip #4
To cooper-atively investiigate a problem					Grp/ Manip #5							
To write a composition												
To explain relationships	Ind/P&P #2	Ind/P&P #3			Grp/Manip #5							

Key: Ind/P&P = Individual responds to paper and pencil task
Self/Manip = Individual uses manipulatives to explore
Grp/Manip = Group completes hands on investigation

Note: Numbers refer to performance assessments (1-5)

APPENDIX B

SAMPLE PERFORMANCE ASSESSMENT ITEM SETS

STUDENT WORKSHEET

Instructions: Read and answer each question below.

1. The sun appears to rise in the ________ and then moves across the sky and sets in the ________.

2. Why does the sun appear to us to move across the sky? (Write your answer in the space below.)

---.

3. If I was travelling from Miami, Florida to somewhere near the North Pole in January, I might find the following:

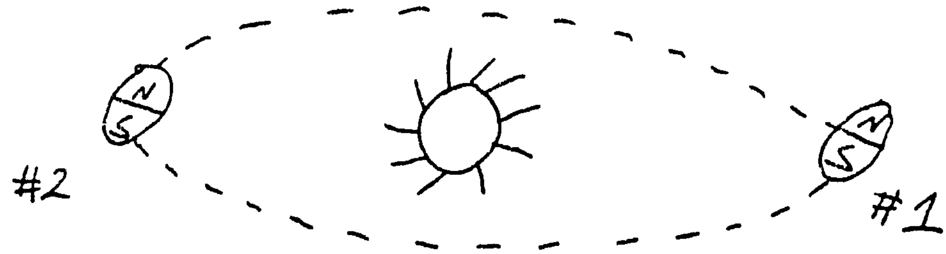
1.) Miami, FL:	Sunrise: 7:00 am Sunset: 6:00 pm
2.) Winston-Salem, NC:	Sunrise: 7:15 am Sunset: 5:15 pm
3.) New York, NY:	Sunrise: 7:30 am Sunset: 4:30 pm
4.) Near North Pole:	Sunrise: 10:30 am Sunset: 1:30 pm

a. Make a chart in the space below showing how many daylight hours each location has. Make sure you label your chart so someone would know what the numbers mean.

b. Which location has the least amount of daylight? ____________________

c. What do you conclude from the data in your chart?

--.



4. The diagram above shows the earth at two points during the year as it revolves around the sun. The northern hemisphere is labelled with a N and the scuthern hemisphere with a S.

a. At point #1 in the above diagram, which hemisphere is having its winter (northern or southern)? ________________

b. At point #1 in the above diagram, which hemisphere is having its summer (northern or southern)? ________________

c. Explain your reason for the answers you gave.

---.

d. Now, draw a diagram of the earth at two points during the year, if we had no temperature or season changes. Use the space below to draw your diagram.

e. Now, explain how your diagram is different from the one at the top of the page.

____

____

____

____.

STUDENT WORKSHEET

Instructions: Read and answer each question below.

1. Explain in the space below what you think causes shadows.

---

2. You have the following information about a shadow.

Journal Entry:

At 9:00 am, the shadow is 6 feet from the base of the tree. At 10:00 am, it is 4 feet away from the base. At 11:00 am, it is 2 feet long. At 12:00 noon, it is 0 feet long. At 1:00 pm, it is 2 feet long. At 2:00 pm, it is 4 feet long. At 3:00 pm, it is 6 feet long. At 4:00 pm, it is 8 feet long.

Make a chart or graph in the space below that summarizes the data from the journal entry. (Make sure you label your chart or graph so someone would know what the numbers mean.)

3. Look at the data in your chart and describe what happens to the length of the shadow between 9 a.m. and 4 p.m.

--.

4. The shadow is shortest when the sun is (circle your answer.):

   a.) to the east of the tree.
   b.) to the west of the tree.
   c.) directly over the tree.

5. At 9:00 a.m. the shadow is west of the tree; at 4:00 p.m. it is east of the tree. Explain in the space below why you think this happens.

--.