DOCUMENT RESUME

ED 358 168 TM 019 970

AUTHOR Rafferty, Eileen A.
TITLE Urban Teachers Rate Maryland's New Performance Assessments.
PUB DATE Apr 93
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Elementary Education; *Elementary School Teachers; Evaluation Methods; Evaluation Problems; Grade 3; Grade 5; Grade 8; State Legislation; *State Programs; *Student Evaluation; Teacher Attitudes; Testing Programs; Thinking Skills; *Urban Schools; Urban Teaching
IDENTIFIERS Alternative Assessment; Baltimore City Public Schools MD; *Maryland School Performance Assessment Program; Open Ended Questions; *Performance Based Evaluation; Teacher Surveys

ABSTRACT

Maryland, the first state to mandate performance assessments for its elementary school students, administers the Maryland School Performance Assessment Program as a week-long series of activities measuring reading, writing, language in use, science, and social studies for students in grades 3, 5, and 8. Open-ended questions are built around activities or tasks that often involve hands-on manipulation. Urban teachers at 144 Baltimore City Public School sites, in Baltimore (Maryland), and school-based staff were surveyed regarding the program, with attention to procedural items (teacher comfort with mechanical aspects of these tests), preparation and administration, and attitudes about the tests. Of 1,436 forms distributed, 404 were returned, a response rate of just over 28 percent. Results show that tests were rated most positively for grade 8, less so for grade 5, and somewhat negatively for grade 3. Teachers expressed concerns about being underprepared for the activities, and concerns about methods of administration. Overall reviews of the tests were equivocal. Although 48 percent rated the tests as not essentially worthwhile, most respondents felt that they measured thinking skills better than multiple-choice tests did, and most felt that the tests would change educational practices. Four tables present survey findings. (SLD)

URBAN TEACHERS RATE MARYLAND'S NEW PERFORMANCE ASSESSMENTS

Eileen A. Rafferty, Baltimore City Public Schools

AERA Annual Meeting Presentation, Spring 1993, Atlanta, GA

Urban Teachers Rate Maryland's New Performance Assessments

Eileen A. Rafferty, Baltimore City Public Schools

AERA Annual Meeting Presentation, Spring 1993, Atlanta, GA

Perhaps more than most innovations to impact the classroom, performance assessments are the product of a new ethic. Differences between performance assessments and traditional multiple-choice tests range from procedural issues, through preparation and administration, to fundamental philosophical shifts. Conflicts, unexpected and often unacknowledged, arise when school personnel apply standards and values ingrained by years of working on multiple choice tests to their judgments of the newer instruments.

Maryland, the first state to mandate performance assessments for its elementary students, has provided the nation a laboratory in which to evaluate the effectiveness of such instruments. The Maryland School Performance Assessment Program (MSPAP) is a week-long series of activities, measuring Reading, Writing, Language in Use, Science and Social Science in which students in grades three, five and eight are tested in randomized classroom-size groupings. Open-ended questions are built around certain activities or tasks which often involve hands-on manipulation of materials (called manipulatives) by small groups of three or four students to solve a problem. Two or, less frequently three, subject areas may be integrated within a given activity. In one imaginative task, fifth graders, in groups of four children conduct an experiment to measure the distance traveled by a model airplane as a function of how many times its propeller is wound.

Following two consecutive spring administrations of the MSPAP, Baltimore's urban teachers and schoolbased staff were surveyed regarding the program. This paper summarizes some of the strengths and problem areas inherent in the performance assessments as observed by the staff who administered them. The final survey focused items on three broad areas: procedural issues, preparation/administration, and judgments/opinions.

Procedural items, Part I, assessed teachers' comfort with mechanical aspects of the new tests. Of particular interest was how school personnel dealt with characteristics unique to the MSPAP such as testing within randomized groupings (instead of everyday classroom groupings), hands-on small-group activities, and limited preview of testing materials.

Preparation/Administration, Part II, compared the frequency of certain classroom practices during regular class time with the frequency of similar practices during the test itself. Preliminary interviews suggested some confusion about what practices were allowed during MSPAP testing. In striking contrast to typical multiple choice tests, for example, the MSPAP encourages the use of dictionaries and thesauruses during the test. Also of interest in this section was the impact of classroom and test-taking practices on attitudes toward the test in general.

Judgments/Opinions, Part III, was designed to survey attitudes about tests and testing which may help or hinder acceptance of performance tests. Issues examined in this section included teachers' views on cooperative group activities, the use of manipulatives in testing, as well as broad conclusions regarding the worth of the instruments and their impact on classroom activities.

Methods

Survey forms were sent to staff located at 144 Baltimore City Public School sites during June 1992. Each school received forms addressed to the principal, the testing coordinator, the alternate testing coordinator, and every teacher/administrator who participated in testing. Participants were guaranteed anonymity with respect to their names and locations.

Respondents represented 95 different testing sites or about 66 percent of all schools. Of the 1436 forms distributed, 404 completed surveys were returned for an individual response rate of just over 28 percent. T-tests, comparing responding and non-responding schools, showed no significant differences in the size of the school or size of testing groups. Although the difference was not statistically significant, test scores at schools in which staff responded to the survey averaged about two normal curve equivalent (NCE) points higher on Total Reading and Total Mathematics sections of the Comprehensive Test of Basic Skills, Fourth Edition than did schools in which staff had not responded.

Participants rated 40 objective items on a five-point Likert scale. The scoring was such that a 1 indicated the most favorable opinion, while a 5 indicated the least favorable opinion. Participants were allowed to choose a "no answer/not applicable" as well as a neutral response. The 40 items were divided into three distinct sections addressing Procedural Issues (Part I), Preparation and Administration (Part II), and Judgment and Opinion (Part III). The sum of a participant's responses was divided by the number of items addressed to yield an average overall rating. Subscale scores for the three different sections were computed in a similar manner. Items to which the participant chose "no answer/not applicable" were not averaged into ratings.

Results

Overall Ratings

Table 1 presents the results of statistical procedures* performed to discern the effect of grade level on test ratings. The analyses, which controlled for the position of the rater (administrator or teacher), average size of testing group, and disposition of special education students (i.e., percent of special education students who had received special testing accommodations and percent of special education students who had been exempted from testing) showed a significant effect of grade level on the Overall Rating, on ratings of Procedural Issues, and on Judgments and Opinions ($p < .01$).

In general, staff members who rated fifth or eighth grade tests reported the most positive attitude toward the MSPAP, while observers of the third grade test reported the least favorable attitude. Some respondents, usually testing coordinators or assistant principals, monitored both third and fifth grade classes. Other raters, employed in schools servicing both third and fifth graders, did not specify a specific grade level. These two populations were aggregated to form

a group designated as "3rd and/or 5th" grade raters. Their ratings tended to be somewhat more positive than those of fifth grade observers.

Overall ratings of the fifth and eighth grade tests garnered positive means of 2.85 and 2.86, respectively. Observers of third grade tests, however, averaged a slightly negative mean of 3.07. The average rating across all grade levels of the test was 2.89.

Table 1. Overall Ratings of General Characteristics Averaged by Grade Level

The average scores below represent a mean based on a scale of from 1 to 5 where lower scores indicate a more favorable response. A score of 3.0 is neutral or uncertain. An average score above 3.0 indicates a response which tended to be unfavorable.

Grade(s) Rated	Number of Raters	Overall Rating*	Part I Procedural Issues*	Part II Preparation & Administration	Part III Judgment & Opinion*
Grade 3 only	105	3.07	3.11	2.74	3.33
Grade 5 only	88	2.85	2.77	2.66	3.16
Grade 8 only	102	2.86	2.67	2.90	3.02
Grades 3/5	103	2.80	2.74	2.58	3.04
Total**	404	2.89	2.83	2.72	3.14

* General Linear Model Maximum Likelihood Regression analyses showed averages for Overall Rating, Part I and Part III differed significantly with the grade level of the test being rated ($p < .01$).
** Includes the responses of 6 participants who could not be categorized by grade level.

Procedural Issues

Part I, Procedural Issues, consisted of items measuring the staff's perception of concerns related to mechanical aspects of the MSPAP administration. Of particular interest was how school personnel dealt with characteristics unique to the MSPAP such as randomized testing groups, hands-on small-group activities and limited access to preview activities. As with Overall Ratings, average ratings on Procedural Issues (Table 1) differed significantly with grade level ($p < .01$). Eighth grade raters were most satisfied with the procedural aspects of the test, averaging 2.67. Somewhat less enthusiastic, but nevertheless positive, were fifth grade observers with a mean of 2.77. The most difficulty with the test's procedural tasks was reported by raters of the third grade test who averaged a somewhat negative rating of 3.11. Not surprisingly, ratings on this section were also significantly related to the number of students in the testing group. All else being equal, raters were more likely to report positive outcomes for procedural issues when the average size of the testing group was small.

Table 2. Part I: Procedural Issues - Item Statistics

Number observations: 404	--- Percent of Participants Responding to Each Alternative ---							
Question	Excellent	Satis-factory	Neutral	Unsatis-factory	Very Unsatis-factory	No Answer	Item Mean	Item Correlation
Were testing group rosters clear?	34	47	10	3	3	2	1.9	.36
Reorg. into random groups go smoothly?	28	45	10	9	4	3	2.1	.39
Ss comfortable working in small groups?	18	47	17	12	3	3	2.3	.51
Did stud.assess. help answer questions?	8	20	17	5	6	45	2.6	.48
+Enough room to perform experiments?	18	39	12	15	14	3	2.7	.53
+Enough room on desk/tables?	14	42	11	18	14	2	2.7	.55
Accommodation guidelines clear?	7	34	19	13	8	20	2.8	.57
+Exemption guidelines clear?	7	31	17	16	9	20	2.9	.48
Rate the prep of examiners at your school	14	28	11	17	24	7	3.1	.47
+Did hands-on activities go smoothly?	8	29	16	26	17	3	3.2	.66
Adequate time for each activity?	8	26	8	24	32	3	3.5	.45
Enough time to study examiners manual?	5	18	7	20	46	4	3.9	.41

+The responses to these items differed significantly with grade level.
All item correlations are statistically significant ($p < .0001$).

Shown in Table 2 is the percentage of participants who responded to each of the sections six possible choices: Excellent, Satisfactory, Neutral, Unsatisfactory, Very Unsatisfactory, and No Answer. (Rounding error may prevent these percentages from summing to exactly 100.) In addition to the percentage of responses for each category, a mean rating was calculated for each item. Note that the items have been sorted by the mean score so that the most positively rated questions appear at the top of the table. Such means have a possible range of from 1.0, excellent, to 5.0, very unsatisfactory. A mean of 3.0 represents a neutral rating. Participants who chose the "No Answer" alternative were not included in the calculations of the mean response.

Perhaps the most favorable result in this section, confirmed that 65 percent of all participants reported the comfort level of students while working in small groups was satisfactory to excellent. Randomized testing group rosters, the reorganization of students into random groups based on the rosters, answering questions about the test by phone and at workshops, exemption guidelines, and guidelines for accommodating special needs students were all rated positively by the participants.

Problem areas reported in this section included limitations on time teachers were given to read the Examiner's Manual, time limitations of the test in general, and the organization of "hands on" activities. The majority of staff felt that the test creators had not allotted adequate time for each activity. The most frequent complaint in this section, often remarked upon in the short answer responses, was the limited amount of time teachers were given to study the *Examiners' Manual* before giving the test. A total of 66 percent of all participants felt that they had not enough time with the materials to prepare for the test. Many added that although they

had access to the *Examiners' Manual*, they were still baffled because most of the instructions and reading materials were not included the *Examiners' Manual*. This information was contained in Student Response and Resource books which were not accessible to teachers before the test.

Other problems areas were marked by significant differences between responses at different grade levels. A majority, 56 percent, of all third grade observers reported that their hands on activities had not gone smoothly. This was in contrast to 48 percent in fifth grade and 36 percent in eighth grade who had reported similar problems. Ratings of space considerations and exemption guidelines also varied significantly with grade level. Concerns about classroom and desk space were more prevalent in the lower grades. About half of all fifth grade raters found the exemption guidelines somewhat unclear, while raters in other grades reported no problem.

The correlation between a raters' responses to an individual item and his/her average on the Overall Rating, based on all 40 items, allows an indication of the relative contribution of that item to the raters' overall satisfaction with the test. In general, respondents were more likely to be satisfied with the test as a whole when the "hands-on" activities went smoothly at their site ($r = .66$). Other responses predictive of overall satisfaction included having adequate space on the students' desk and in the classroom to perform experiments, understanding the guidelines for special education accommodations, and reporting that students were comfortable working within small groups.

Preparation and Administration

The summary statistics (Table 1) for this section reveal that average responses on Preparation and Administration were positive. In fact, unlike other portions of the survey, grade level was not a significant predictor of composite scores on Preparation and Administration. Participants who observed the third and/or fifth grades showed the most positive score at 2.58. Fifth grade raters posted a mean of 2.66, followed by third grade raters at 2.74. Observers of eighth grade tests scored 2.90.

Table 3. Part II: Preparation and Administration - Item Statistics

Number observations: 404	– Percent of Participants Responding to Each Alternative –							
Question: Do students...	Frequently	Often	Some-times	Rarely	Never	No Answer	Item Mean	Item Correlation
+work in small groups during class?	39	30	21	4	0	6	1.9	.35
+use manipulatives during regular class?	32	28	26	5	1	7	2.1	.32
do hands-on activities during regular class?	26	30	30	6	1	6	2.2	.35
use a dictionary/thesaurus during class?	25	31	25	5	2	11	2.2	.29
consult maps during class?	19	32	25	7	2	15	2.3	.31
use graphic organizers during class?	15	26	29	8	4	19	2.5	.38
use MSPAP tools/equipment during class?	16	24	32	12	4	12	2.6	.36
consult maps during test?	6	19	35	17	8	15	3.0	.40
+use dictionary/thesaurus during test?	10	23	24	18	15	10	3.0	.42
use graphic organizers during test?	5	15	32	25	10	13	3.2	.42
perform MSPAP-like science experiments in class?	9	13	26	24	10	19	3.2	.41
practice in randomized test groups before test?	14	15	21	9	29	12	3.3	.35
work in small groups w/in randomized test groups?	10	14	20	12	33	11	3.5	.40

+The responses to these items differed significantly with grade level
Correlations over .30 are statistically significant ($p < .0001$). Those below .30 are significant at $p < .001$.

Displayed in Table 3 are item statistics, sorted in order of mean response, on the 13 questions which comprised the Preparation and Administration section. The majority of respondents, regardless of grade level, indicated that practices such as working in small groups, using tools and manipulatives, performing hands-on activities, consulting dictionaries, thesauruses and maps, and constructing graphic organizers for writing exercises were events which took place often or frequently during regular classroom hours.

Participants reported that certain practices were more frequently employed during regular classtime than during the test itself. Fifty-six percent of all respondents reported that their students used dictionaries and thesauruses frequently or often during regular class time. However, just one-third believed that students were consulting these references frequently or often during testing. A majority of respondents, 51 percent, indicated that maps were frequently or often consulted during regular classtime, whereas just 19 percent said that maps were consulted with similar frequency during the test. This occurred in spite of the fact that several of the tests included exercises which required map reading. Approximately 41 percent of respondents cited use of graphic organizers during classroom writing activities, whereas just 20 percent observed their use with the same frequency during testing. Of course, some of the reported discrepancies may be attributable to the difficulty encountered when rating the frequency of behaviors within very different time spans. It is also probable, that some staff members failed to recognize and promote the use of reference materials during MSPAP testing.

Of all the items in this section which assess regular classroom practices, only the frequency of MSPAP-like science experiments was rated less than optimally. Just 22 percent of respondents said such exercises occurred often or frequently, while 34 percent reported that the experiments were rarely or never performed in class.

Staff felt that some students might be inhibited by the change of classmates and (occasionally) teachers which resulted from the MSPAP mandate to test children in randomized testing groups rather than in their usual class setting. Students were also required to form into randomly assigned cadres of three or four to work on particular exercises. In most cases, students had never before worked together within the specific randomized groupings required by the test.

Responses to some items were found to be significantly related to grade level. The classroom practices of working in small groups and using manipulatives was reported to occur more frequently in the third and fifth grades than in the eighth grade. Eighth graders, however, were more likely to consult a dictionary or thesaurus during the test when compared to lower classmen.

Judgments and Opinions

In contrast to performance assessments, multiple choice tests emphasize individual tasks and reward students who have outperformed their classmates. Cooperative behavior among students can at best be thought to invalidate results and at worst is considered cheating. By contrast, the MSPAP requires that students work in cooperative groups to solve problems. The test rewards cooperative behavior rather than individual effort.

In addition to cooperative-competitive conflicts, group work affects classroom climate. Successful group activities are often noisy. The quiet and orderly work once considered the hallmark of scholarly pursuits is disrupted. Rather than all children leaning from a single adult, children learn and ask questions of one another. To the extent that teachers have been taught to value order and individual achievement, they may have difficulty understanding and approving of the newer tests.

Among the responses to Part I, Procedural Issues, was a single response which was most predictive of overall satisfaction with the test. Staff who reported that their hands-on activities went smoothly were more likely to hold a favorable attitude toward the test in general. Such activities are inevitably small group projects in which students manipulated objects and performed experiments. One respondent characterized one of the tasks as follows, "We were examining the soil sample. Some kids spilled water. Some made mud pies." Attitudes explored in this section of the survey were intended to reflect the value systems used in judging the test.

Two items in this section specifically addressed feelings about group work. Fifty-five percent of all participants disagreed or strongly disagreed with the statement that, "Group activity is too noisy and disruptive." Thirty-two percent agreed or strongly agreed that, "When students work in groups, the scores are not fair." An additional two items measured participants' thoughts on the use of manipulatives. When asked whether the use of tools and manipulatives should be discontinued, 66 percent said no. Fifty-three percent approved of the statement that, "Manipulatives help focus the students' attention." Other items focused on the contrast between the open-ended, problem-solving assessments typical of performance tests and the more standardized, content knowledge assessments usually seen in multiple choice exams. Thirty-six percent of the participants disagreed with the statement, "MSPAP directions should be open-ended." However, 59 percent agreed that tests such as the MSPAP measure thinking skills better than do multiple choice exams."

Also addressed were notions related to concepts of locus of control. These attempted to gage the staff's perception of their ability to influence the test's outcome. Only 5 percent of respondents disagreed or strongly disagreed with the item, "MSPAP scores can be improved by using certain classroom practices." Although most participants felt that, by using certain classroom practices they could improve scores, a substantial number, 33 percent, felt that MSPAP would have little or no effect on classroom practices. When asked if current classroom practices had changed for the better because of the MSPAP, 41 percent said no. Nevertheless, 76 percent were of the opinion that current classroom practices were not adequate to prepare students for the MSPAP.

Table 4. Part III: Judgments and Opinions - Item Statistics

Number of Observations: 404 — Percent of Participants Responding to Each Alternative —

Questions	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree	No Answer	Item Mean	Item Correlation
Scores can improve w/ certain classroom practices	32	42	15	5	5	2	2.1	.34
Discontinue the use of manipulatives/tools	11	9	20	42	16	2	2.4	.46*
MSPAP measures thinking better than multiple choice	24	35	19	11	11	2	2.5	.41
Group activity is too noisy & disruptive	11	19	12	43	12	2	2.7	.52*
Manipulatives help focus attention	13	40	21	15	9	2	2.7	.48
MSPAP will have little effect on classroom practices	18	15	24	29	13	1	2.9	.51*
When Students work in groups, scores are not fair	14	18	27	29	10	2	3.0	.45*
Students were interested in the test materials	6	35	20	21	16	2	3.1	.43
+Most students finished before time was up	12	26	11	26	23	2	3.2	.18
Classroom practices are better because of MSPAP	5	16	36	21	20	2	3.3	.52
MSPAP directions should be open-ended	5	12	35	24	22	3	3.5	.13
+MSPAP is essentially worthwhile	6	18	25	19	29	3	3.5	.52
+Students would have done better if given more time	31	17	25	18	8	1	3.6	.19*
Current practices are adequate to prepare for MSPAP	3	7	14	34	42	1	4.1	.38
Some tasks were too difficult for students	63	23	5	2	4	1	4.5	.27*

* For these correlations the number of points awarded was reversed to ensure that low scores consistently represented favorable outcomes.
+ The responses to these items differed significantly with grade level
Item correlations greater than .30 were statistically significant ($p < .0001$). Correlations below .30 were significant at the $p < .001$ level.

Discussion and Recommendations

The results presented in the preceding section show that the tests were rated most positively in the eighth grade, less positively in fifth grade, and somewhat negatively in third grade. Mean statistics show that most procedural/clerical requirements, including testing children within randomly assigned groups, were handled successfully. Aspects rated unfavorably on the procedural/clerical scale reflected teachers' concerns about being under-prepared for the activities. Asked to categorize the frequency with which certain strategies were used during regular instruction vs. during test administration, teachers most often responded that students consulted maps and dictionaries, worked in small groups, used manipulatives and graphic organizers more often in the classroom than on the test itself. Sample assessments may help prepare both teachers and students for the unfamiliar formats characteristic of MSPAP tasks. In addition, testing coordinators should ensure that schooltime be set aside for teachers to review MSPAP materials in the week before the test begins.

Both short answer items and responses to judgmental statements regarding MSPAP indicated areas that must be addressed if the assessments are to be accepted by school personnel. A small but significant number of staff members, for example, felt that group activities were too noisy and disruptive and that test scores would not be fair so long as students were allowed to work

in groups. Others felt that the use of manipulatives should be discontinued. An analysis of short answer responses was helpful in isolating misconceptions arising from prior experience with standardized testing practices. As an example, an activity encouraged on the MSPAP, consulting a dictionary or thesaurus during the test, was considered cheating by some teachers. Others indicated that the test would have no effect on how children were taught in their schools because "It is absurd to believe that a test should drive curriculum." These results suggest that state and local administrators can help redirect teachers' perceptions of the scientific rationale of the test by offering inservice workshops specially targeted at addressing such misconceptions.

Overall reviews of the test were equivocal. Despite averages that were slightly positive, negative means were observed for third grade observers. Negative responses were also characteristic of scores on the judgments and opinions section of the survey. Although 48 percent responded that MSPAP was not "essentially worthwhile," nearly all felt that it measured thinking skills better than multiple choice tests did and indicated that classroom practices would change as a result of the test. Whereas most respondents felt that the MSPAP had not as yet effected desirable changes in classroom practices, others were adamant in the opinion that current teaching strategies were not adequate to prepare students for the assessments. Insofar as there is motivation within the school to raise test scores, we can expect that such classroom practices will begin to approximate those which will best prepare students for the test.