DOCUMENT RESUME

ED 400 291 TM 025 562

AUTHOR	Myerberg, N. James
TITLE	Inter-rater Reliability on Various Types of Assessments Scored by School District Staff.
PUB DATE	Apr 96
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Accountability; Achievement Tests; Educational Assessment; Elementary Secondary Education; *Evaluators; *Interrater Reliability; Language Arts; Mathematics Tests; *School Districts; Scores; *Scoring; Test Format; Test Use; *Training
IDENTIFIERS	High Stakes Tests; Monitoring; *Montgomery County Public Schools MD; Short Answer Tests

ABSTRACT

The Montgomery County (Maryland) public school system has started using assessments other than multiple-choice tests because it is felt that this will provide school staff with better information about the success of the instructional program. One of the ways assessments can provide better information is by having teachers score student papers. This, however, can conflict with another goal of the assessment program, high-stakes accountability for schools. An immediate solution to this potential source of conflict has been to have teachers score the papers in a centralized setting with extensive training and control, including the random assignment of papers. The three tests that the system has scored in this way are mathematics short-answer, mathematics extended-answer, and language arts extended-answer tests. Scorers had intensive training and close monitoring. Scoring consistency was evaluated by correlations between scorers and the percent of large differences between scorers. Reliability results indicate that constant, active monitoring is required to achieve consistent scoring, and that it is more difficult to score language arts assessments consistently than mathematics assessments. Attachments present the mathematics scoring rubric and a sample scoring report a rater would receive. (Contains three tables.) (SLD)

Inter-rater Reliability on Various Types of Assessments Scored by School District Staff

N. James Myerberg
Montgomery County (Md.) Public Schools

Paper presented at the annual meeting of
The American Educational Research Association in April 1996
in New York

Inter-rater Reliability on Various Types of Assessments Scored by School District Staff

The Montgomery County (MD) Public Schools have started using non-multiple choice assessments because it is felt this will provide school staff with improved information about the success of the instructional program. One of the primary ways in which these assessments can provide better information is by teachers scoring student papers. This way the teachers can see the type of work that students are doing. However, teacher scoring of papers can conflict with another goal of the assessment program -- high stakes accountability for schools. Pursuit of these conflicting goals is certainly not unique to Montgomery County. This paper is intended to show how we are dealing with this conflict and to provide other districts with comparative data. Our program is still developing and we are interested in similar information from other districts who are scoring their own non-multiple choice assessments.

The immediate, and possibly intermediate, solution to the conflicting goals stated above was to have a group of teachers score the papers in a centralized setting with extensive training and control. This way any favoritism toward papers from their school could be monitored. Most of the papers that the teachers scored were randomly assigned to them from all the schools in the district. This provided them with an opportunity to see the quality of work throughout the district and gave them a new perspective on how their own students performed. However, they also were given up to 50 papers from their own school so they could see the level of performance of those students.

Description of Tests

The three tests that were scored are described below.

Math short answer contained 10 questions, each scored from 0 to 3 points and was administered in Grades 3 to 8. The assessments used in Grades 4 and 6 were the Mathematics Goals Tests developed by the Psychological Corporation. The assessments in the other grades were locally developed.

Math extended answer was one multi-step activity scored holistically from 0 to 6 points. It was administered in Grades 4, 6 , and 7 and was locally developed.

Language arts extended answer consisted of one reading and writing activity scored from 0 to 4 points for each of 3 domains -- Response to Reading, Management of Content, and Command of Language. It was administered in Grades 4, 6, and 7. It is part of the Language Arts Performance Assessment series developed by the Psychological Corporation.

Data Collection

There were 99,232 test papers scored by 169 teachers in 13 days. Broken down by the 3 types of tests; there were 50,019 math short answer tests, 24,710 math extended answer tests, and 24,503 language arts extended answer tests. The math tests were scored by 101 teachers and the language arts tests by 68 teachers.

The original plan was to have each paper scored twice to achieve the best reliability for each student's score. However, budget constraints would not permit this. The revised plan was to double score at least 50 percent of the language arts papers and 25 percent of the math papers so we would know if we were getting acceptable inter-rater consistency. The actual numbers and percents of papers double scored on each test are shown in the attached tables. We eventually double scored 81 percent of the language arts papers and 26 percent of the math papers. Part of the reason for the difference from expectation on the language arts was, as we monitored the scoring we found we were being less consistent with the language arts. Another reason was we overestimated the time teachers would take to score the language arts papers. Since the scorers were promised 13 days of work, they kept scoring even though we had already completed what we planned to do.

Training and Monitoring of Scorers

The scorers had intensive training before they started scoring and close monitoring and help, if needed, while they were scoring. The initial training was done by workshop leaders. However, most of the monitoring and help was provided by group leaders who had been selected for that job because of previous scoring experience and teaching and curriculum development expertise. The scorers worked in groups of 10 to 12 for language arts and 15 to 18 for math. The several steps in the training and monitoring are described below.

1. There was a general discussion of scoring and the specific instrument to be scored. The discussion emphasized the importance of consistent scoring across scorers. A major point made to the scorers was that the papers should be scored according to the rubrics established in the workshop, not how the teacher would score the paper in her or his classroom. We spent time on this point because it had been a problem area in scoring the previous year.

2. The scorers took the test they would be scoring to familiarize themselves with it. This was followed by more discussion of how the instrument would be scored. This varied by the nature of the instruments as described below.

   For the math short answer, each item was broken down into the specific parts that needed to be evaluated. The number of these that were correct was then related to the 0 to 3 scale. For example, if an item had 4 parts to be evaluated, all 4 correct earned 3 points; 2 or 3 correct earned 2 points; and 1 correct earned 1 point. Attachment A shows the relationship between the number of correct parts and point earned.

For the math extended answer, the various steps in the activity were identified as major and minor. The number of each of these that was incorrect was then related to the 0 to 6 scale. For example, a mistake on a major step meant the student could not receive a score of 6.

For the language arts extended answer, the scoring was based on a more general rubric for each of the three domains. These rubrics were developed by the Psychological Corporation. As an example, part of the statement for 4 points from the Command of Language rubric is "Sentences are correctly written and they display variety. Expository responses exhibit clear and precise word choices."

3. A common set of 5 papers was scored and discussed. These papers had previously been scored by an expert group made up of a workshop leader and the group leaders. The discussion began with each person announcing the score they gave the paper. Those who gave the paper a score other than that agreed to by the "experts" were asked to give a reason for their score. This led to an exchange of ideas that helped the scorers to clarify what was expected. These discussions were often spirited and even led to a couple cases of the "experts" changing their scores.

4. Step 3 was repeated with 5 more papers. This second set almost always produced better agreement among scorers.

After Step 4 the scorers were placed in their groups and began scoring independently. Monitoring the scoring was quite important because poor scorers could not be dismissed. This was a summer in-service activity for teachers and they were promised a specific number of days of work. The monitoring took the following 3 forms.

5. As the scoring began, the group leaders would also review 1 or 2 of the papers for each scorer to see if additional training was needed.

6. Group leaders were available for consulting whenever a scorer found a student answer that they felt had not been covered in the training. Sometimes the group leader could relate the answer to something in the training. At other times the answer was indeed one that had not been seen or anticipated previously. When this happened the group leaders would meet with the workshop leader to determine how the response should be scored. The results of these meetings would be shared with all scorers of that test. While this was sometimes disrupting, it was needed to maintain consistency across multiple groups of scorers.

7. Scorers received reports of how they were scoring. Attachment B is a sample of these reports. Consistency of scoring was emphasized in the data given to the scorers. The reports showed the mean scores that each scorer gave to a set of papers. This mean was compared to the mean given to those same papers by a random sample of other scorers. These reports quickly became very popular with the scorers since they provided an easy way for them to see how they were doing. Many of the scorers who had originally shown a tendency to score high or low anxiously awaited the next round of reports to see if they had improved and in most cases they had.

Results

The scoring consistency was evaluated using two measures -- correlations between scorers and the percent of large differences between scorers. Large differences were defined as greater than 1 point on all measures except the math short answer total test. On that 30 point score a large difference was anything greater than 4 points.

Math short answer -- The correlations between scorers on the 60 individual items (scored from 0 to 3) across Grades 3 to 8 ranged from .72 to .97 with the median being .88. The percent of large differences ranged from 0 on 5 items to 11 on 1 item. The median large difference was 2 percent.

The correlations on the total test (scores ranged from 0 to 30) were .96 in 2 grades and .97 in the other 4 grades. Four of the grades had 2 percent large differences and the other 2 grades had 3 percent large differences. Data related to the math short answer assessments are presented in Table 1.

Math extended answer -- The correlations between scorers on the 7 point (0 to 6) activities across the 3 grades tested ranged from .88 to .90. Large differences ranged from 4 to 5 percent. Data related to the math extended answer assessments are presented in Table 2.

Language arts extended answer -- The correlations between scorers on the 9 domains (scored 0 to 4) across the 3 grades tested ranged from .54 to .69 with the median at .57. The percent of large differences ranged from 2 to 8 with the median at 5. Data related to the language arts extended answer assessments are presented in Table 3.

Comparison of results from different tests -- Each of the assessments discussed in this paper had a different number of score points. Therefore, to compare the correlations from the various assessments, the correlations have been adjusted using the Spearman-Brown formula.

The consistency of scoring was better for math than for language arts. This was true for both types of math assessment and the short answer items as well as the total test.

- o For the math short answer items, the correlations were substantially higher than for language arts even though the math had fewer score points (4 vs. 5).

- o For the math extended answer, the correlations were higher than for the language arts, even after the correlations were adjusted. The highest language arts correlation, .69, would be .77 after being adjusted.

- o Comparing the language arts to the math short answer total test shows that the adjusted language arts correlation ranged from .90 to .94, still less than the .96 and .97 for the math. However, this is merely a theoretical comparison because no non-multiple choice activity would be scored on one 30 point scale.

It is difficult to compare the results between the math short answer items and the math extended answer assessment because the latter has only 3 correlations compared to 60 for the short answer. However, after adjusting the short answer correlations to account for the difference in score points, it appears the short answer items were generally scored with more consistency. The median correlation for the short answer items, .88, would be .94 if adjusted for the difference in score points. This is higher than the extended answer correlations. The highest extended answer correlation, .90, would be .82 if adjusted for score points. This is lower than 46 of the 60 short answer correlations

When the math short answer total test score is compared to the math extended answer, the results favor the latter. The correlations for the math extended answer assessment are slightly higher than for the math short answer total test score. The adjusted correlations for the extended answer assessments would be .97 and .98; the original correlations for the short answer total test were .96 and .97. As with the language arts assessment, this is merely a theoretical comparison.

Discussion

Two of the main conclusions we reached from the activities described in this paper are:

- o It is more difficult to consistently score language arts assessments than math assessments, and

- o Constant, active monitoring is required to achieve consistent scoring.

Certainly neither of these is new or surprising. However, they are important points to emphasize for any school district that is undertaking scoring their own non-multiple choice assessments, especially if the results will be high stakes.

The consequence of the first point above is that when scoring a language arts assessment, especially if poor scorers cannot be easily dismissed, each paper must be scored twice. While tight budgets might make this difficult, our data show that there can be considerable error in individual scores if double scoring is not done in a situation in which poor scorers must be allowed to continue. While this would certainly be advisable to have with math assessments also, our data show it might not be necessary. However, even in that case, some double scoring is needed to get a good estimate of the level of consistency. This would probably be at least 100 papers per scorer as long as those papers are randomly assigned to a second scorer.

The discussion of monitoring the scorers earlier in this paper emphasized the importance of that activity. This point is raised again here because it could become a major issue in my district and possibly in other districts. There is a push in Montgomery County to move the scoring into the schools. This certainly makes monitoring more difficult.

Moving the scoring into schools is being pushed for at least 2 reasons. One is that the teachers will see the students' work and be better able to plan their instructional programs. While this represents one of the most valuable aspects of non-multiple choice assessments, we also found there was value in teachers scoring the papers of other students in the district. They were able to gain some perspective on how their students were doing by seeing what other students were doing.

The second reason for moving to scoring in the schools is tight budgets. If this scoring can be made part of the teachers' regular job then no money is need for a summer scoring workshop. However, until it can be shown that scoring in the schools will produce consistent, unbiased scoring, these centralized, intensely monitored workshops are needed. They are certainly needed when school scoring is started to establish a baseline of what the quality of scoring should be.

Attachment A

MATHEMATICS

SCORING RUBRIC FOR OPEN-ENDED QUESTIONS

(Point values associated with part responses)

NUMBER OF PROBLEM PARTS	POINT VALUES FOR NUMBER OF PARTS CORRECT		
	3 POINTS	2 POINTS	1 POINT
8	7,8	4,5,6	1,2,3
7	6,7	3,4,5	1,2
6	5,6	3,4	1,2
5	4,5	2,3	1
4	4	2,3	1
3	3	2 or 1 with major parts of others	1 or parts of others
2	2	1 or part of 2	part of 1
1	1	part	attempt

Attachment B

```
NUMBER OF VALID OBSERVATIONS (LISTWISE) =        85.00

VARIABLE       MEAN     STO OEV    MINIMUM    MAXIMUM VALID N  LABEL

TOTA         15.824      7.423       1.00      29.00      85   REAOER SCORES
TOTB         15.494      7.174       3.00      29.00      85   OTHER REAOERS SCORES
OIFF           .329      2.296      -4.00      10.00      85   MEAN OIFFERENCE BETWEEN REAOERS
AOIFF         1.671      1.599        .00      10.00      85   ABSOLUTE OIFFERENCES
OIFF1          .082       .277        .00       1.00      85   PROPORTION 3 OR MORE POINTS BELOW
OIFF2          .259       .441        .00       1.00      85   PROPORTION 1 OR 2 POINTS BELOW
OIFF3          .224       .419        .00       1.00      85   PROPORTION SAME SCORE
OIFF4          .318       .468        .00       1.00      85   PROPORTION 1 OR 2 POINTS GREATER
OIFF5          .118       .324        .00       1.00      85   PROPORTION 3 OR MORE POINTS GREATER
OIFFI1         .047       .596      -1.00       3.00      85   MEAN OIFFERENCE - ITEM 1
OIFFI2         .094       .548      -1.00       2.00      85   MEAN OIFFERENCE - ITEM 2
OIFFI3        -.035       .606      -3.00       3.00      85   MEAN OIFFERENCE - ITEM 3
OIFFI4         .259       .774      -2.00       3.00      85   MEAN OIFFERENCE - ITEM 4
OIFFI5         .000       .000        .00        .00      85   MEAN OIFFERENCE - ITEM 5
OIFFI6        -.176       .640      -3.00       1.00      85   MEAN OIFFERENCE - ITEM 6
OIFFI7         .106       .535      -1.00       2.00      85   MEAN OIFFERENCE - ITEM 7
OIFFI8         .024       .617      -3.00       3.00      85   MEAN OIFFERENCE - ITEM 8
OIFFI9         .000       .408      -1.00       2.00      85   MEAN OIFFERENCE - ITEM 9
OIFFI10        .071       .530      -2.00       3.00      85   MEAN OIFFERENCE - ITEM 10
```

Table 1
Scoring Quality Data for Math Short Answer, Grade 3, 1995
(N=2528,%=30)

Item	Percent Same Score	Percent Adjacent Score*	Percent Large Differences**	Inter-Rater Reliability
1	64	33	3	81
2	85	11	4	82
3	66	31	4	80
4	86	13	2	92
5	68	29	2	84
6	94	6	0	97
7	95	5	0	97
8	77	22	1	81
9	76	21	4	84
10	67	31	2	82
Total	25	72	3	96

Scoring Quality Data for Math Short Answer, Grade 4, 1995
(N=2525,%=29)

Item	Percent Same Score	Percent Adjacent Score*	Percent Large Differences**	Inter-Rater Reliability
1	82	16	2	86
2	80	17	4	91
3	86	12	2	87
4	83	16	2	88
5	85	14	1	86
6	66	28	5	81
7	79	18	2	87
8	69	26	6	81
9	91	7	2	93
10	84	10	5	89
Total	28	70	2	96

* - Adjacent scores are differences of 1 point for items and 1 to 4 points for the 30 point total test.

** - Large differences are 2 or more points for items and 5 or more points for the 30 point total test.

Table 1 (continued)
Scoring Quality Data for Math Short Answer, Grade 5, 1995
(N=2787,%=33)

Item	Percent Same Score	Percent Adjacent Score*	Percent Large Differences**	Inter-Rater Reliability
1	74	22	4	86
2	86	13	2	88
3	80	18	2	90
4	83	16	2	92
5	81	16	3	91
6	84	15	1	93
7	73	24	4	82
8	81	15	4	90
9	79	21	0	88
10	85	13	1	94
Total	27	71	2	97

Scoring Quality Data for Math Short Answer, Grade 6, 1995
(N=1502,%=18)

Item	Percent Same Score	Percent Adjacent Score*	Percent Large Differences**	Inter-Rater Reliability
1	95	4	1	85
2	80	16	4	87
3	92	6	2	95
4	93	5	2	88
5	92	8	0	96
6	73	25	2	86
7	87	11	2	92
8	73	20	7	81
9	71	25	4	81
10	78	20	2	89
Total	32	66	2	97

* - Adjacent scores are differences of 1 point for items and 1 to 4 points for the 30 point total test.

** - Large differences are 2 or more points for items and 5 or more points for the 30 point total test.

Table 1 (continued)
Scoring Quality Data for Math Short Answer, Grade 7, 1995
(N=2486,%=30)

Item	Percent Same Score	Percent Adjacent Score*	Percent Large Differences**	Inter-Rater Reliability
1	84	15	1	93
2	88	10	1	92
3	71	22	7	81
4	70	23	7	82
5	84	13	2	90
6	68	27	5	72
7	86	12	2	93
8	84	9	7	87
9	91	7	1	96
10	87	12	1	94
Total	29	68	3	97

Scoring Quality Data for Math Short Answer, Grade 8, 1995
(N=1678,%=22)

Item	Percent Same Score	Percent Adjacent Score*	Percent Large Differences**	Inter-Rater Reliability
1	87	11	1	95
2	83	14	3	86
3	83	15	2	91
4	71	18	11	74
5	82	16	2	72
6	79	17	3	86
7	83	14	3	82
8	91	9	0	95
9	82	16	1	89
10	76	20	3	84
Total	29	69	2	97

* - Adjacent scores are differences of 1 point for items and 1 to 4 points for the 30 point total test.

** - Large differences are 2 or more points for items and 5 or more points for the 30 point total test.

Table 2

Scoring Quality Data for Math Extended Answer, 1995

Grade	Percent Same Score	Percent Adjacent Score	Percent Difference GE 2	Inter-Rater Reliability	Number	Percent
4	60	36	5	88	2645	31
6	56	38	5	90	1692	21
7	61	35	4	90	1881	23

Table 3

Scoring Quality Data for Language Arts Extended Answer, 1995

Grade	Domain	Percent Same Score	Percent Adjacent Score	Percent Difference GE 2	Inter-Rater Reliability	Number	Percent
4	Response to Reading	59	40	2	67	7361	86
	Management of Content	58	40	2	69		
	Command of Language	59	39	2	64		
6	Response to Reading	54	40	6	57	6693	83
	Management of Content	49	44	7	55		
	Command of Language	51	43	6	55		
7	Response to Reading	46	47	8	54	5884	75
	Management of Content	49	46	5	57		
	Command of Language	51	45	4	61		

AERA April 8-12, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Interrater Reliability on Various Types of Assessments Scored by School District Staff.

Author(s): N. James Myerberg

Corporate Source:

Publication Date: April 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ Sample sticker to be affixed to document

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document ☐

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: N. James Myerberg	Position: Consultant
Printed Name: N. James Myerberg	Organization: Montgomery County (MD) Pub. Sch.
Address: 5606 Ogden Rd. Bethesda, MD 20816	Telephone Number: (301) 229-9447
	Date: April 16, 1996