DOCUMENT RESUME

ED 347 198 TM 018 712

AUTHOR McLaughlin, Donald H.
TITLE Effects of Administrator Performance on Student Performance in the Trial State Assessment.
PUB DATE Apr 92
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administrator Role; *Administrators; Grade 8; Junior High Schools; *Junior High School Students; *National Surveys; *Performance; Scores; Student Evaluation; Testing Problems; Test Results; *Test Use; Test Validity; Training
IDENTIFIERS *Examiner Effect; Monitoring; National Assessment of Educational Progress; Test Directors; Testing Conditions; *Trial State Assessment (NAEP)

ABSTRACT

In 1990 the scope of the National Assessment of Educational Progress (NAEP) was broadened by adding the Trial State Assessment (TSA), in which approximately 2,500 eighth graders in 100 schools in 40 states and U.S. territories participated in the mathematics assessment. A major step was training local test administrators to administer the TSA sessions. Members of the NAEP contractor's professional staff were assigned to monitor half of the TSA sessions. A critical question for TSA validity was whether students' performance would differ for monitored and unmonitored sessions. The competence and objectivity of local administrators were issues of great importance. There were small, but reliable differences in the effect of some aspects of the testing environment on the performance of students. When students were cooperative and administrators were proficient, an ideal situation was created for students to demonstrate mathematics proficiency. The differences found were small, but steps should be taken to ensure that the same ideal testing situation is available for all students. Observations of training of TSA administrators confirmed expectations of the highly professional quality of that training. Nevertheless, continued monitoring of state NAEP sessions seems warranted to ensure uniform testing conditions. Nine tables present study data. (SLD)

Effects of Administrator Performance on Student Performance in the Trial State Assessment

Donald H. McLaughlin
American Institutes for Research

American Educational Research Association
April 23, 1992

Background

In 1990, the scope of the National Assessment of Educational Progress (NAEP) was significantly broadened by adding the Trial State Assessment (TSA), in which a sample of approximately 2,500 eighth graders in 100 schools in each of 40 states and territories participated in the mathematics assessment. A major methodological step was taken by training local test administrators to administer the TSA sessions, in contrast to the professional testing staff who administer Mother NAEP. This was a bold methodological step because there was great concern that, on the one hand, local school staff might be tempted to make things easier for their students, and on the other hand, these staff had only one day's training in the administration of NAEP.

Therefore, members of the NAEP contractor's professional staff were assigned to monitor half of the sessions. The selection of sessions for monitoring was random, and local administrators did not know until minutes before the session whether it was to be monitored. A critical question for the validity of the TSA was whether students' performance would differ between monitored and unmonitored sessions. Furthermore, where student performance is affected by attributes of the testing situation, including

the performance of the local test administrator, inferences about the performance of these students are problematic.

Results

The main result is favorable for the validity of State NAEP, as shown in Table 1. Before proceeding, I must explain the entries in Table 1. The left-hand columns of numbers (the "basic measure") are measured in units that have an overall standard deviation, over 100,000 cases, of one[1]. The right-hand columns of numbers were obtained by subtracting from each student's score the mean score of students in the same sex by race/ethnicity group. Therefore, the adjusted scores contain absolutely no main effect of race/ethnicity or sex.

INSERT TABLE 1 ABOUT HERE.

The unit of analysis here is the session; and a high scoring session, in the left-hand columns, is a session in which the average of students' scores was higher than the average over all sessions, while a high scoring session in the right-hand columns is a session in which the members of each group (e.g., white boys) score higher than the average for that group over all sessions. Comparisons of the two types of statistics, basic and adjusted, is useful because it indicates whether effects of particular variations in the testing conditions are felt equally by all students or are greater for one group or another.

---

[1] The basic measure is the logit transform (log(p/(1-p)) of the percent of items the student answered correctly, of those reached. It is corrected for differences in the items each student answered by setting the mean to zero and the variance to one separately for the students who were exposed to each booklet.

The first two rows of numbers in Table 1 are the means for the 1,918 monitored and 1,997 unmonitored sessions. That is, the difference in performance between monitored and unmonitored sessions was less than two percent of a population standard deviation. The next three rows of Table 1 compare the variance components of scores in monitored and unmonitored sessions. The variance components are estimated directly, by solving normal equations. The major result here is that there was no substantial tendency for greater variation among unmonitored sessions than among monitored sessions. Of some interest, however, is the comparison of variance components between columns: clearly, if the sessions had been composed of homogeneous race/ethnic and sex groups, variation between sessions within states would have been reduced, and variation between states would have been reduced even more.

Finally, the figures in the bottom row of Table 1 provide indications of the likelihood of finding a mean difference this large by chance. These "t-values" are based on the variance components presented. They do not make use of the case weights, however, which I must explain. In this particular paper, I am not presenting results as representative of the population of American schools nor as representative of schools in any particular state: the questions I am addressing are methodological in nature (effects of variations in administration of testing on results of testing), and as such, the most powerful test is obtained by weighting each student equally. This being said, the "t-values" at the bottom of Table 1, both less than 1.96 in absolute value, suggest that it would not have been unusual to have obtained these differences between monitored and unmonitored sessions by chance.

Although the performance scores were very similar across the entire mathematics assessment, might there have been greater differences if we focused on the parts of the assessment individually (numerical operations, measurement, geometry, data analysis, and algebra)? The results in Table 2 suggest that any differences are very slight, only about one point on each of the NAEP subscales. A difference of one

on the NAEP zero-to-five-hundred scale represents less than two percent of the eighth grade standard deviation. That is, the results presented in Table 2 match those in Table 1 and do not depend on the choice of statistic. Results based on the logit percent correct, adjusted for booklet differences, are the same as results based on the IRT-based plausible values for the underlying skills.

INSERT TABLE 2 ABOUT HERE.

There is a reason to expect that student performance in the monitored sessions might average slightly higher, if administration problems should occur, and it cannot be denied that they occur! In monitored sessions, the monitor, a highly trained testing professional, is empowered to intervene when a significant problem arises. For example, if the administrator became confused while presenting examples of use of calculators, the monitor could help to clarify the instructions for the students. Therefore, it is of interest to find out what kinds of sessions, if any, had greatest differences between monitored and unmonitored status. The results in Table 3 suggest that differences were larger in schools with large percentages of minority students enrolled. Although the differences in performance between monitored and unmonitored sessions in schools with mostly white students was only about one percent of a standard deviation, the differences in other schools was about four percent of a standard deviation, enough of a difference to be unlikely to occur by chance.

INSERT TABLE 3 ABOUT HERE.

It is important to note that this is not a function of the race/ethnic group of the student whose performance is being measured: the differences were as large for performance measures from which any race/ethnic variation has been eliminated as for "raw" performance measures. That is, in (some) schools in which large percentages of students were minorities, NAEP participants of all race/ethnic groups were likely to obtain lower scores than others in their same race/ethnic group in other schools.

What is the mechanism of this difference? It could be either the local administrator's performance or the environment created by the other students in the assessment session. First, to verify that local administrators' performance might be related to student performance, one can compare student performance between sessions in which administrators were rated as more or less effective. In monitored sessions, monitors rated the overall performance of the administrators on a one-to-five scale. A rating of "one" was given when no assistance was needed, "three" for mixed results, and "five" if it could not have been done without the monitor's presence. Dividing the sessions into those with ratings of one and two, versus all others, yields the results in Table 4. There is a substantial difference, as might be expected. Students in sessions in which the administrator received a poor rating scored .13 standard deviations lower than students in other sessions. On the NAEP scale, that is a difference of about five points.

INSERT TABLE 4 ABOUT HERE.

Perhaps the most interesting aspect of Table 4, however, is the finding that for the adjusted scores the difference was less than half as large. The conclusions to be reached from this are either (a) that the affects of poor administrator performance (as

rated by the monitor) were felt primarily in sessions with large proportions of lower scoring groups, (b) that poor administrator performance occurred more frequently in sessions with more lower scoring groups, or (c) both. Examination of results in Table 5 suggest that both factors may have operated, although the relation between administrator performance ratings and the ethnic minority status of the school enrollment was not statistically significant (chi square=3.07, 1 d.f.). The differences between basic and adjusted measures occurred in both categories of schools: relations between administrator ratings and student performance were stronger in the measure from which race/ethnic and sex variance had not been removed.

INSERT TABLE 5 ABOUT HERE.

Why should there be this differential effect? To address this, consider the session from the perspective of the administrator. In both monitored and unmonitored sessions, administrators were asked to rate how well the session went. In nearly 90 percent of both monitored and unmonitored sessions, administrators responded "very well" to this question. The results in Table 6 suggest that in sessions not marked "very well": student performance was lower, by more than .18 standard deviations, in both monitored and unmonitored sessions. Results in Table 6 also suggest a particular locus of these sessions where "how it went" was related to lower performance: unmonitored sessions with large proportions of minority participants, in schools with large percentages of minority students. In these 62 sessions, performance was more than one quarter standard deviation lower when the administrator reported that things did not go very well.

INSERT TABLE 6 ABOUT HERE.

Did "not going very well" refer to the level of student cooperation or to the local administrator's performance? To address this, we can look separately at the monitors' ratings of (a) student cooperation and (b) administrator performance of particular tasks. In about seven percent of the sessions, the monitor indicated that students were not cooperative and orderly during the session; and as shown in Table 7, performance in these sessions was about .18 standard deviations lower than in other sessions. There were, of course, no comparable monitor ratings for the unmonitored sessions.

INSERT TABLE 7 ABOUT HERE.

Monitors recorded ratings of dozens of items describing the sessions, and in evaluating the validity of the 1990 Trial State Assessment, a panel of the National Academy of Education looked at all of them. Factors such as the time of day, adequacy of the room, and occurrence of timing errors bore no significant relation to average session performance.

Two types of deviation from nominal procedures that might affect performance were errors in instructing students on the use of hand-held calculators and errors in instructing students on the use of the test booklets. The results in Tables 8 and 9 suggest that in sessions in which monitors noted deviations from nominal administrator

performance, student performance was lower. For calculator training, the difference was small, about .05 standard deviation, and not statistically significant; and this difference disappeared when race/ethnic and sex effects were removed. However, for test booklet instructions, the difference was .16 standard deviations. We expected that the effects of calculator instructions might be more noticeable for numerical operations, measurement, and data analysis, than for geometry or algebra, but there was no substantial variation among those subscales.

INSERT TABLES 8 AND 9 ABOUT HERE.

Conclusion

The overall conclusion is that there were small, but reliable differences in the effect of some aspects of the testing environment on the performance of students participating in a session. When students were cooperative and administrators were proficient, an ideal situation was created for each student to demonstrate his or her proficiency in mathematics. When these conditions were not met, measured performance was lower, and there are indications that this decrement in performance was most noticeable in measures from which variation related to race/ethnicity and gender had not been removed. While the differences described here were very small in relation to variation of performance in the population of eighth graders, additional steps should be taken to ensure that the same ideal testing situation is available for all participants. Perhaps, given expectations that testing sessions in some schools may be more problematic than in others, especially highly qualified administrators might be selected for those schools.

Observations of the training of local administrators by the NAEP contractor confirmed our expectations of the highly professional quality of that training. Nevertheless, continued monitoring of State NAEP sessions seems warranted, as well as efforts to ensure that students of all types are tested in uniform conditions.

Table 1. Performance Measures for Monitored and Unmonitored Sessions

Standardized Logit Scores	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Monitored Sessions (n = 1918)	Unmonitored Sessions (n = 1997)	Monitored Sessions	Unmonitored Sessions
Mean	.009	-.009	.010	-.010
Within-Session Variance	.803	.800	.766	.768
Within-State Between-Session Variance	.137	.144	.076	.080
Between-State Variance	.087	.091	.031	.033
t-value for Mean Difference	1.24		1.86	

Table 2. Mean Mathematics Sub-Scales in Monitored and Unmonitored Sessions

Sub-Scale	Monitored Sessions	Unmonitored Sessions
Numerical Operations	266.1	265.3
Measurement	258.4	257.7
Geometry	259.9	258.8
Data Analysis and Statistics	261.6	260.6
Algebra and Functions	260.9	259.8
Composite	262.0	261.1

Table 3. **Performance Measures**, by School Minority Membership Percentage, for Monitored and Unmonitored Sessions

Standardized Logit Scores	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Monitored Sessions	Unmonitored Sessions	Monitored Sessions	Unmonitored Sessions
Schools with Large Minority Percentage	-.313 (n=584)	-.357 (n=598)	-.072	-.115
t-value for Difference	1.73		2.20	
Schools with Small Minority Percentage	.148 (n=1334)	.139 (n=1390)	.046	.034
t-value for Difference	0.60		0.88	

Table 4. **Mean Student Performance**, by Monitor Ratings of Overall Administrator Performance

Measure	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Good (1-2) (n=1614)	Poor (3-5) (n=251)	Good (1-2)	Poor (3-5)
Logit Percent Correct	.027	-.107	.017	-.032
t-value for Difference	3.23		1.73	
Numerical Operations	266.8	262.0		
Measurement	259.2	253.1		
Geometry	260.6	255.4		
Data Analysis/Statistics	262.6	255.8		
Algebra and Functions	261.6	256.5		
Composite	262.7	257.3		

Table 5. Performance Measures, by School Minority Membership Percentage, for Sessions with High and Low Ratings

Standardized Logit Scores	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Good (1-2)	Poor (3-5)	Good (1-2)	Poor (3-5)
Schools with Large Minority Percentage	-.296 (n=481)	-.400 (n=89)	-.068	-.095
t-value for Difference	2.04		0.72	
Schools with Small Minority Percentage	.162 (n=1133)	.049 (n=162)	.052	.001
t-value for Difference	2.25		1.34	

Table 6. Performance Measures, by "How It Went," for Monitored and Unmonitored Sessions

Standardized Logit Scores	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Monitored Sessions	Unmonitored Sessions	Monitored Sessions	Unmonitored Sessions
Sessions that went Very Well	.030	.012	.027	.003
Other Sessions	-.138	-.178	-.101	-.116
t-value for Difference	3.81	3.12	4.21	3.25
Schools with Large Minority Percentage				
Sessions that went Very Well	-.294 (n=495)	-.326 (n=536)	-.057	-.101
Other Sessions	-.421 (n=89)	-.623 (n=62)	-.159	-.230
t-value for Difference	2.17	2.96	2.57	2.30
Schools with Small Minority Percentage				
Sessions that went Very Well	.166 (n=1176)	.155 (n=1238)	.062	.047
Other Sessions	.014 (n=158)	.004 (n=152)	-.070	-.070
t-value for Difference	2.95	2.82	3.32	2.937

Table 7. Mean Student Performance, by Monitor Ratings of Student Cooperation and Orderliness

Measure	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Good (1) (n=1731)	Poor (2-3) (n=134)	Good (1)	Poor (2-3)
Logit Percent Correct	.024	-.181	.019	-.099
t-value for Difference	4.08		3.36	
Numerical Operations	266.7	258.8		
Measurement	259.1	249.5		
Geometry	260.4	252.3		
Data Analysis/Statistics	262.4	251.5		
Algebra and Functions	261.5	253.4		
Composite	262.6	254.0		

Table 8. Mean Student Performance, by Monitor Ratings of Calculator Training

Measure	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Good (1) (n=1657)	Poor (2-3) (n=233)	Good (1)	Poor (2-3)
Logit Percent Correct	.015	-.034	.009	.018
t-value for Difference	1.21		-0.30	
Numerical Operations	266.4	264.1		
Measurement	258.7	256.0		
Geometry	260.1	258.3		
Data Analysis/Statistics	262.1	258.5		
Algebra and Functions	261.2	258.7		
Composite	262.3	259.8		

Table 9. Mean Student Performance, by Monitor Ratings of Script Errors in Test Booklet Instructions

Measure	Basic Measure		Adjusted for Race/Ethnic and Sex Differences	
	Good (1)	Poor (2-3)	Good (1)	Poor (2-3)
Logit Percent Correct	.040	-.119	.027	-.061
t-value for Difference	3.96		3.23	
Numerical Operations	267.2	261.9		
Measurement	259.7	253.2		
Geometry	261.0	255.1		
Data Analysis/Statistics	263.2	255.5		
Algebra and Functions	262.0	256.2		
Composite	263.2	257.1		