Report on Wyoming's Testing Irregularities

Richard Hill
Center for Assessment

July 27, 2010
Revised: July 30, 2010
Revised August 27, 2010

Background

Wyoming has administered the multiple-choice portion of its Proficiency Assessments for Wyoming Students (PAWS) on-line since 2006 (the open-ended questions are administered with paper and pencil). In 2009, a new contractor (Pearson) was hired, and Pearson ran the first year of the testing program using the platform that was in place when they took over the contract. In 2010, however, Pearson used a new platform. Many administration problems were reported, both to the Wyoming Department of Education (WDE) and Pearson. Problems included long waits, lost work requiring students to restart test from beginning, and students being incorrectly identified as not having taken the practice test.

There was widespread concern throughout the state that the administration problems had affected student performance. There was much discussion within the state about concerns over impact of these administration problems on test scores. As a result, the WDE decided to delay the reporting of test results until the scope of the problem was better understood and recommendations could be made about what an appropriate response would be.

WDE hired the Center for Assessment to study this issue and to (1) determine likely impact of administration problems and (2) make recommendations relative to reporting of results: At what level(s) should reports be produced, and with what caveats?

Initial Efforts

The Center began the contract by recognizing that there were two major areas of investigation to be conducted:

1. Documentation of the problem: How often did problems occur, what was the nature of the problem(s), and to whom did they happen? Did the problem(s) occur for individual students, or for classes, schools or districts? Did the problem(s) occur randomly or systematically?
2. What was the impact on achievement when the problem(s) occurred? The essence of the plan was to identify high impact vs. low impact groups, and then using prior year's achievement as a covariate, attempt to isolate effect size.

A major concern here was to not confound the two issues by trying to explore both at the same time. If, for example, an attempt was made to identify the impact on all the students affected at the same time, it was likely that many unaffected students would be inadvertently included in the analysis, thereby diluting the effect. Therefore, the goal in the second area of investigation was to find groups of students who were unquestionably affected by administration issues and those that were clearly not, even if these were relatively small samples of those groups. If we could determine the impact of

the problem when it happened, that information would be useful in trying to determine how often it happened. As a result, we decided to tackle the second area first.

As we started to identify groups that had been affected and not affected, it was clear that some validation of those groups would be critical. If the information we had caused us to mislabel students or groups, the validity of the entire study would be brought into question. So, rather than inferring which students had been affected and presuming that those inferences were accurate, an important step in the study would be to have local school people double-check those lists. So, the general plan for proceeding was this: Pearson would attempt to create a list of students (or classes or schools) who likely had been directly affected by the administration issues, and WDE staff would confirm the accuracy of those lists with local school staff. At the same time, WDE would try to independently come up with its own list of affected students or groups by directly contacting local school staff and asking them to provide information about affected students or groups. After that, we would see how those students performed this year relative to their performance last year. Whatever decline we saw in their relative performance this year would be attributed to the administration issues, and that would be a first step in trying to resolve what reports should be produced this year.

Another possible way of identifying the impact of the administration problems was to take advantage of the fact that students took the open-response questions on paper, and therefore their answers to these questions were unaffected by the on-line administration problems. Their scores on these questions could serve as a covariate, similar to that of prior year's achievement.

Identifying an Affected Group

Pearson maintains a toll-free call center for every administration of the test. Even in a year when testing goes smoothly, the call center gets numerous calls, generally related to asking for information about testing procedures or asking for needed materials. Every call is logged on what is referred to as a "ticket." Pearson has a procedure for ensuring there is appropriate follow-up to all tickets. When the administration problems occurred this year, many calls, in addition to the usual volume, came in from local school staff to report the problems and to ask to direction on what to do as a result. These tickets seemed to be a logical place to start to identify a group of students who had been affected.

This year, there were 1,549 tickets created as a result of calls. Of these, 489 described some problem with the on-line administration; the remaining 1,000+ were routine calls about other issues. Pearson staff placed the 489 tickets into one of three categories:

a. An issue with a specific student was identified
b. An issue with a specific small group of students was identified
c. The issue did not identify a specific student or group

Combining the first two categories, Pearson identified about 400 students, across all grades and subjects, who clearly had been affected by administration problems. At the same time, WDE was finding it was having problems generating its list. Contemporaneous logs often did not have information that would permit them to identify the specifics of who had been affected, and attempts to contact local school personnel often were unsuccessful because school people had left for the summer and could not be reached. That same issue made it impossible to validate Pearson's list of 400 students, but the documentation associated with those students was so strong that it was deemed worthwhile to proceed with that list without validating it.

Data and Results

The study was limited to reading and mathematics; it was presumed that any effect found there would carry over to the other content areas. Also, it was limited to grades 4-8, since prior year's achievement would not be available for students in grades 3 or 11. Students who did not have data in the previous year also had to be eliminated. For all these reasons, the number of students available for study was reduced from the 400 mentioned previously to under 200; 119 in reading and 53 in mathematics. However, if the administration problems had had a substantial impact on student achievement, it should be evident from a group of even this limited size, so we decided it was worthwhile to proceed with the analysis of the data. In addition, there was confidence that these students had been unquestionably affected by administration problems.

Pearson computed the deviation of these students' scaled scores (divided by the standard deviation of student scaled scores, so the results would be reported in a standardized form) from the state average in 2009 and 2010. Table 1 provides several statistics for each grade; the number of students included in the study, the standardized deviation of those students' performance from the state mean in both years, and then the difference between those results, the paired student-level standard deviation, the paired t-test, and the probability of that paired t under the null hypothesis of no change in deviation between the years. To increase the power of the study, results also are totaled across all grades.

Table 1

Test Results in Reading and Mathematics
For Students Identified as Affected by Administration Problems from Pearson's Tickets

Content Area	Grade in 2010	N	Deviation in 2009	Deviation in 2010	2010 Deviation – 2009 Deviation			
					Mean	SD	t	P(t)
Reading	4	36	-0.030	-0.011	0.019	0.775	0.147	88.37%
	5	19	0.266	-0.306	-0.572	0.674	-3.701	0.16%
	6	9	-0.494	-0.481	0.013	0.394	0.101	92.23%
	7	32	-0.431	-0.541	-0.110	0.658	-0.951	34.92%
	8	23	-0.434	-0.207	0.227	0.945	1.154	26.10%
	All	119	-0.204	-0.274	-0.070	0.774	-0.991	32.35%
Math	4	17	0.261	0.343	0.082	0.505	0.669	51.33%
	5	3	0.221	-0.206	-0.426	1.051	-0.702	55.53%
	6	9	0.202	0.499	0.297	0.721	1.234	25.23%
	7	8	-0.917	-1.176	-0.259	0.239	-3.068	1.81%
	8	16	-0.438	0.005	0.443	0.642	2.755	1.47%
	All	53	-0.140	0.007	0.147	0.636	1.683	9.83%

Table 1 tells us that the impact of the administration issues on these students' achievement was, at best, minimal. Performance in reading declined, but even with an N of 119, the decline was so small that it was not statistically significant. Performance in math *increased*, but again, the change in performance was so small as to not be statistically significant. As an additional check, we computed the correlation of students' performance across the years, under the hypothesis that if the impact had affected students differentially, we would find the correlation of performance across years to be less than the range of .70 - .80 that is typical when there are no administration problems. The correlation

for reading across all 119 students was .72; for mathematics, is was .78. In summary, this study provided no evidence that the administration issues had contributed to a decline in student achievement.

Impact on Statewide Performance

It can never be known for sure whether changes in statewide performance could be attributed to administration problems, since results across years might (and do) go up or down for a myriad of reasons. If, for example, the statewide averages declined between 2009 and 2010, that result might be due to problems with the administration, a changing population of students, or a real decline in student achievement. Nonetheless, it seemed reasonable to compare the performance of students across years to see what the changes had been. An examination of the p-values of the equating items across years suggested that statewide performance not only had not declined, but had increased from 2009. As a result, the contractor was asked to equate the scores across years and compute the mean scaled scores.

Table 2 provides the mean scaled scores for 2009 and 2010 for all grades tested in reading, mathematics and science. The means are based on the full populations both years (6500-6800 per grade for grades 3-8, and over 8,000 for grade 11 reading and math, and a little under 6,000 for grade 11 science). All these N-counts are similar across the two years, except for grade 11 reading and math, where the N-counts are 500-700 higher this year than last. This tells us that there likely was little to no change in who was included in the scores across the two years, and thus provides confidence that the mean scores across the years are comparable.

Table 2

Statewide Mean Scaled Scores for 2009 and 2010

Grade	Reading		Mathematics		Science	
	2009	2010	2009	2010	2009	2010
3	585.0	591.7	647.7	649.7		
4	659.6	663.1	655.4	660.2	668.0	664.4
5	654.1	656.4	680.1	679.9		
6	680.9	677.6	706.0	702.8		
7	674.7	674.4	716.5	717.2		
8	693.0	696.0	726.1	726.8	646.8	646.4
11	158.9	163.3	149.2	149.1	154.2	153.7

As can be seen from Table 2, the mean scaled scores are higher in 2010 than in 2009 for a majority of the cells. The exceptions are grades 6 and 7 for reading, grades 5, 6 and 11 for math, and all grades for science. Several of these declines are trivial (well less than 1 scaled score point). The only drops of more than one scaled score point are grade 6 reading and math, and grade 4 science. Without an explanation of how the administration problems could have affected grade 6 without affecting the other grades, one must assume that the decline in scores at that grade was due to reasons other than administration difficulties.

Conclusions

We were limited in the studies we could do because of logistical issues, but we were able to look at two sets of data that should have shed light on the impact of the administration problems. The first study, which looked at a limited number of students who were reported to have had problems with administration, showed that those students scored as well, relative to the state average, in 2010 as they had in 2009. The second study simply looked at the statewide averages in 2010 and compared them to the averages for 2009. In both years, the averages included all students, and the N-counts across the years suggest that the two tested groups were equivalent. While scores at some grades were down, the average change was positive—even with the administration problems, students scored higher, on average, in 2010 than they had in 2009. So neither study provided evidence that the administration problems had a negative impact on student performance.

That does not mean, of course, that no students were affected, or even that a more controlled study would have not found an effect. But it does mean that if there was an effect, it was limited, both in its scope and its impact on student performance.

We therefore make the following recommendations:

1. All reports that were originally planned should be produced and distributed. Without evidence to the contrary, it should be assumed that the reports provide a valid estimate of student achievement.
2. If it is known that a student was affected by administration problems, and the achievement of the student on PAWS was inconsistent with other information about the student, the PAWS result likely should be discarded. Note, however, that this recommendation is consistent with all good testing practice; any time an individual test result is not consistent with other known information about a student's achievement level, the other information should take higher priority in judging the student.
3. WDE and Pearson should make an offer to any district that feels it can identify subgroups that were clearly affected and clearly not affected to conduct the kind of impact study we were unable to do under the time constraints provided by this contract.