



Use of Statewide Assessment Results: How to Use Results to Inform Instruction and How the Results Should NOT be Used



Jon S. Twing, Ph. D.
Executive Vice President - Pearson



Overview of Presentation

What information do we get from assessments when they work and how best to use it?

What impact was discovered following the technical problems with the spring administration?

How should the PAWS results be interpreted in light of this technical impact?

What implications does this impact have for longitudinal and trend analyses?



What is Psychometrics?

Definition

- Learning Theory + Measurement + Statistics

Origins

- Psychology and IQ testing

Mathematical Underpinnings

- Classical Test Theory
- Item Response Theory (IRT)

Famous Psychometricians...



What is Assessment?

General Definition

- Act or result of judging the worth or value of someone or something
 - Rate
 - Amount
 - Size...

Educational Definition

- Process by which information is obtained relative to some known objective or goal

Testing and assessment usually used as synonyms but can mean different things



Critical Components of Assessments

Validity

- Is this test measuring what it is intended to measure?
- Types of validity evidence include:
 - Content
 - Concurrent / predictive
 - Construct
 - Consequential



Reliability

- How consistently does this test measure?
 - Internal consistency
 - Test/re-test, alternate forms



Spring PAWS Testing Challenges

Technical problems with Pearson's online platform in 2010

Many administrative problems reported

- Long Waits
- Testing interruptions
- Lost responses upon restart
- Other



Widespread concern that administration problems affected student performance



Guidance from WDE

Consulted with their technical advisory committee

Conducted Webinars to solicit stakeholder input

Petitioned Feds to waive use of PAWS 2010 results for accountability purposes

Engaged with a external evaluator to consider what uses, if any, could be made of the spring 2010 PAWS results



External Evaluator

Dr. Richard Hill from the Center for Assessment

Was hired to:

- Determine likely impact of administration problems and
- Make recommendations relative to reporting of results
 - Should reports be produced and if so at what level of aggregation?
 - What caveats, if any, are needed for the score interpretations resulting from the reports?





Focus of Evaluation

First step in the evaluation was to find students who likely had been directly affected by the administration issues

Second step was to examine the performance of the identified students compared with other students, taking their 2009 PAWS scores into account as a covariate

- A covariate was needed to help untangle the confounding between real achievement and impact from the technical problems



Identifying Affected Students

Pearson searched through “tickets” logged by call center during the administration

1,547 tickets created as a result of calls

489 referred specifically to administration problems. These fell into three categories:

- An issue with a specific student
- An issue with a specific small set of students
- An issue that did not specify a specific student or group



Identifying Affected Students

Combining the two categories of specifically identified students resulted in about 400 students across all grades and subjects

This list was further reduced to about 200 students taking either reading and mathematics in 2010 and for whom 2009 PAWS results were also available

While these numbers are low, as the consultant pointed out in his report, the power is being able to compare those we are certain were impacted with those we suspect were not



Score Comparisons

Scores for the affected students in 2009 and 2010 were turned into “deviation” scores (i.e., they were standardized relative to the state mean and standard deviation)

The mean and SD of the difference between 2010 and 2009 was calculated

Statistical tests (dependent t-tests) were carried out on these differences



Affected Student Score Comparisons

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%



Interpretations

Little evidence can be seen suggesting that the administration issues affected student achievement

- Performance in reading across all grades declined slightly, but was not statistically significant
- Performance in math across all grades increased slightly, but was not statistically significant

Correlations of scores between 2009 and 2010 for the affected students were high

- 0.72 for Reading
- 0.78 for Math



Statewide Score Comparisons

Based on near-final data files, comparisons were made between the statewide mean scale scores for 2010 and 2009 using all students by grade and subject

Results suggested 2010 scale scores were higher than or equal to 2009 scale scores in most cases

Notable exceptions were grade 6 Reading and Math and grade 4 science



Statewide Mean Scale Score Comparisons

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	<u>679.9</u>		
6	680.9	677.6	706.0	702.8		
7	674.7	<u>674.4</u>	716.5	717.2		
8	693.0	696.0	726.1	726.8	646.8	<u>646.4</u>
11	158.9	163.3	149.2	<u>149.1</u>	154.2	<u>153.7</u>



Evaluator's Conclusions

PAWS reports originally planned for 2010 should be produced and distributed

The analyses don't say that any specific individual could not have been affected by the administration problems

Individual results should be interpreted in light of other known information about a student's achievement level

- Match your expectation with the results seen
- If expectations are not met, evaluation all the information



Closing Comments

The data seem to suggest that, despite the administration problems that occurred in spring 2010, students still took the PAWS seriously and gave their best efforts

Although no effects were detected at aggregated levels, effects still could have occurred for individual students

The PAWS score reports provide useful information about student performance, but should be interpreted cautiously

The students of the State of Wyoming were able to show they learned despite the challenges they faced during administration