

PREDICTING MATHEMATICS SCORES

Jonathan R. Brown, Clarion University
Courtney L. Brown, Kent State University

The research problem is that there is ongoing and heated debate about the usefulness of state mandated assessment instruments launched by NCLB (No Child Left Behind, 2004) for reforming public school educational academic practices. Proponents of measurement-driven reform have argued that “if you test it, they will teach it” and that assessment drives the educational system to be more productive and effective (Popham, 1987). A rationale to research the effectiveness of a state mandated test is based on opponents of measurement-driven reform asserting that high-stakes testing and assessment creates negative side effects such as dumbing-down the curriculum, de-skilling teachers, pushing students out of school, and generally inciting fear and anxiety among both students and educators (Darling-Hammond & Wise, 1985; Gilman & Reynolds, 1991; Jones & Whifford, 1997; Madaus, 1988a, 1988b; Shepard, 1989). According to the opponents of measurement-driven reform, these negative side effects outweigh any possible benefits of mandated measurement-driven reform. For high-stakes testing to be practical for school districts, there must be reliable and valid ways to predict how students will perform on these mandated tests. By making reliable and valid predictors available to school districts, curricular changes may be implemented and data-driven decision processes implemented.

Teachers are changing what and how they teach in response to state testing programs as revealed by preliminary results from a multi-state survey. Those changes are greatest in states where more severe consequences are attached to test results (i.e. Pennsylvania), according to the two-year study by researchers at Boston College’s National Board on Educational Testing and Public Policy (Olson, 2002). A higher percent of teachers in high-stakes testing states reported that classroom instruction values changed. Forty-three percent of teachers in high-stakes testing states said that instruction and testing changed “a great deal,” compared with 17 percent in states with moderate-stakes testing for schools and low-stakes testing

for students (Olson, 2002). The question is, however, are the changes that teachers are reportedly making based on reliable and valid assessment information or are the changes being made simply based on loosely built hunches?

Past research studies have not always supported measurement driven educational reforms. Past research has frequently emphasized the negative consequences of high-stakes testing when factors other than test results are examined. For example in the 1980s and early 1990s, high-stakes testing created pressures that encouraged teachers to place unprecedented emphasis on drill-based instruction, narrowing of content, and the regurgitation of facts (Corbett & Wilson, 1991; Smith, 1991). In addition, substantial teaching time was lost in test preparation activities (i.e., learning the test formats rather than additional content).

The rewards to schools for achieving high-test scores on state mandated tests have included incentives such as cash awards. They also have included consequences for schools, individual teachers, and students. These consequences include public reporting of test results to media outlets resulting in critical commentary, prevention of grade-to-grade promotion, prevention of high school graduation, and takeovers of schools that have demonstrated low-test scores. These incentives and consequences are all based on one thing, the reported scores from completion of standardized testing. An important question, however, is do the test score results influence instructional practices in positive ways?

In 1999, Pennsylvania adopted academic standards for reading, writing, speaking and listening, and mathematics. These standards identified what a student should know and be able to do at varying grade levels. School districts were given the freedom to design curriculum and instruction to ensure that students meet or exceed the standards’ expectations. The annual Pennsylvania System of School Assessment (PSSA) was designed as a standards-based criterion-referenced

assessment used to measure a student's attainment of the academic standards while also determining the degree to which school programs enable students to attain proficiency of the standards. Every Pennsylvania student in 5th, 8th and 11th grade is assessed in reading, mathematics, and writing.

Since the PSSA is a state mandated test with the outcome of the test results being used to classify schools as passing or failing, to provide funding to schools, and to label and punish schools with low scores, then a need exists to find reliable and valid predictors of PSSA performance. The PSSA is reported to be reliable and valid (HumRRO, 2004). The PSSA is reported to be correlated positively (HumBRO, 2004) with the SAT (Scholastic Aptitude Test), the CTBS/Terra Nova (Comprehensive Test of Basic Skills), the CAT-5 (California Assessment Test, version 5), the NWEA (Northwest Evaluation Association) tests, and the NSRE (New Standards Reference Exam).

The advantage of having reliable and valid predictors of PSSA performance is that these predictors provide school districts with viable indicators that may be used to align curriculum and instructional practices with the expected learning outcomes to be measured by the PSSA. One frequently given test that is used as a predictor is the Comprehensive Test of Basic Skills (CTBS). The CTBS is designed to measure achievement in reading, language, spelling, mathematics, study skills, science, and social studies. Since the CTBS is reported to be positively and highly correlated with the PSSA (HumRRO, 2004), then it is a predictor that needs to be investigated.

Therefore, the first hypothesis was that the scores students earned on the CTBS (CTBS: Comprehensive Test of Basic Skills) mathematics test given in 2nd grade (total score percentile rank) would be a statistically significant predictor of how the same students performed on the PSSA (Pennsylvania System of School Assessment) mathematics in 11th grade (percentile rank mathematics). The second hypothesis was that, since the school system identified students in 2nd grade that performed poorly in mathematics and subsequently used this information to make curricular and teaching changes to help these low-performing 2nd grade students, the 2nd grade students would have

statistically improved their mathematical ranking as demonstrated by the PSSA mathematics test scores in 11th grade.

Methods

Prior to the investigation, the Institutional Review Board of a Pennsylvania university approved this research study. Participants in the study included twenty K – 12 public school districts in northwestern Pennsylvania. From the pool of twenty school districts, one school district was randomly selected.

The school district participating in this study was given both written and verbal rationale for the study, hypotheses, purpose, procedures, and informed about the disclosure process of the study. The superintendent, principals, and school board of the school district participating in the study approved the research methods. The school board and administrators, students, and parents/caregivers were given assurances that the identity of the school, administration, parents/caregivers, and students that participated in the study would be known only to the school district and principal investigators of this study. All participants were assured that there were no known risks for participating in the study. Information for contacting the principal investigators of the study was given, and contact information for securing the results of the completed study was also provided.

Seventy-five students were eligible to participate in the study and all students participated. The students represented a population characterized as 50.1% female and 49.9 males. Ethnic representation was: 75% Caucasian, 12% African American, 10% Hispanic/Latino, and 3% Asian.

The instrumentation used in this study was the CTBS test and subsequent scores measured by the school district when students were in 2nd grade and PSSA (Pennsylvania System of School Assessment) mathematics test and subsequent scores for the same students during 11th grade. Both tests were reported to be reliable and valid for the populations sampled (HumRRO, 2004). Materials used in the study included: CTBS scores from 2nd grade and PSSA scores for the same students in 11th grade. Scores were

electronically transferred to an Excel spreadsheet and imported into a commercially available statistical package for analyzing the data. Coordination was made between the researchers and the school district's chief administrator to collect test scores as reported in percentile rank for each student (CTBS total percentile rank score for mathematics, PSSA percentile rank mathematics), tabulate scores, and prepare information for analysis.

Results

Descriptive statistics for the CTBS (percentile rank total) and PSSA (percentile rank) are listed in Table 1. The first operation performed with the data was to determine if the distribution of scores for CTBS and PSSA were normally distributed. An Anderson-Darling (Weisstein, 2005) test indicated that both sets of scores were normally distributed. The second operation was to determine if the CTBS scores were correlated with the PSSA scores. Pearson correlation results were that a statistically significant positive correlation of 0.682 ($p < 0.001$) existed.

Table 1

Variable	N	Mean	Standard Error of Mean	Standard Deviation	Minimum Score	First Quartile	Median	Third Quartile	Maximum Score
1PctM	75	58.72	3.16	27.40	6.00	32.00	63.00	81.00	99.00
2 Tot_1	75	78.01	2.32	20.10	14.00	68.00	84.00	96.00	99.00

The third operation was to create a fitted line plot of the 2nd grade percentile rank CTBS mathematics scores (Tot_1) with the 11th grade percentile rank PSSA mathematics scores (PctM). Linear, quadratic, and cubic models were explored. The quadratic model demonstrated an R^2 value of 48.2% (Figure 1). The 11th grade PSSA mathematics scores were predictable and statistically significant ($p < 0.001$) using the quadratic equation:

$$\text{PctM} = 14.42 - 0.0391 \text{ Tot}_1 + 0.0073202 (\text{Tot}_1)^2$$

Therefore, the first hypothesis that the scores students earned on the CTBS mathematics test given in 2nd grade would be a statistically significant predictor of how the same students performed on the 11th grade PSSA mathematics test was supported.

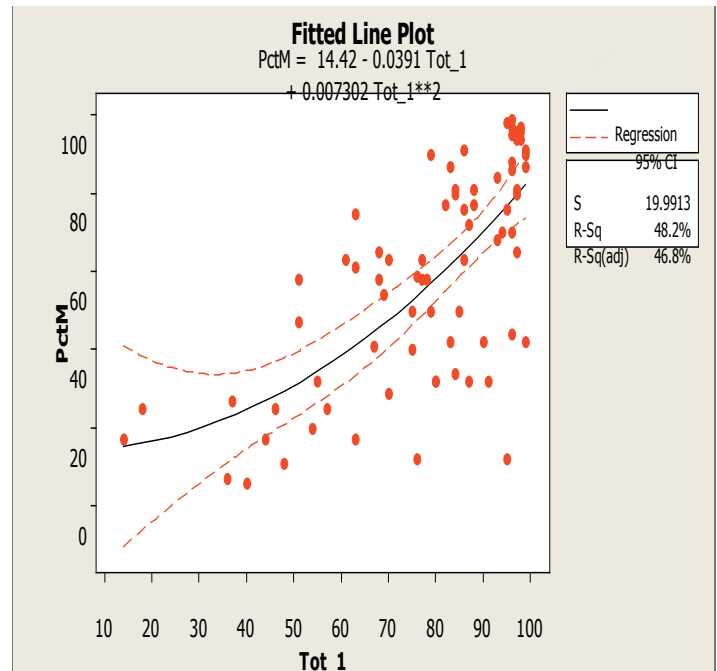


Figure 1: Fit of the quadratic model.

To investigate the second hypothesis, students who performed poorly on the 11th grade PSSA were identified (at-risk students). Performing poorly was defined as any student scoring in the basic or below

basic ordinal classification designated by the Pennsylvania Department of Education (other classifications are: proficient and advanced). Basic and below basic have been operationally defined as scoring at or below the 35 percentile ranking on the PSSA. The Pennsylvania Department of Education has designated that students must perform at the proficient or advanced level on the PSSA to meet the demands of No Child Left Behind (2004). Performing at the basic or below basic level mandates remediation for students and sanctions for school districts.

The first operation was to identify the at-risk student scores by creating a subset from all the original scores. This subset consisted of twenty-four of the original seventy-five students. This subset of scores included all students that scored at or below the 35th

percentile on the PSSA mathematics test and each student's accompanying 2nd grade CTBS score. The second operation performed was to determine if the subset distribution of scores for CTBS and PSSA were normally distributed. An Anderson-Darling (Weisstein, 2005) test indicated that both sets of scores were not normal distributed.

The third operation was to plot the subset of scores for the low-performing students (Figure 2), examine them visually, and determine if the CTBS scores appeared to be correlated with the PSSA scores. A visual inspection indicated that there appeared to be a poor relationship between the two sets of scores.

Therefore, the second hypothesis was that since the school system identified students in 2nd grade that performed poorly in mathematics on the CTBS and subsequently used this information to make teaching and curricular changes to help these low-performing 2nd grade students improve mathematically, that the 2nd grade students would have statistically improved as demonstrated by the PSSA mathematics test scores in 11th grade. Descriptive statistics for this high-risk (low-performing) group is reported in Table 2.

Table 2

Variable	N	Mean	Standard Error of Mean	Standard Deviation	Minimum Score	First Quartile	Median	Third Quartile	Maximum Score
1 PctM<35	24	20.21	2.01	9.86	3.00	14.25	20.00	27.00	38.00
2 Tot_1A	24	62.50	4.67	22.90	14.00	44.50	69.50	80.00	99.00

Figure 2 is a fitted line plot of the data for the at-risk group; visually, it did not support the second hypothesis. The curvilinear fitted line plot did not demonstrate that low or high CTBS scores in 2nd grade were clear predictors of 11th grade PSSA scores. Additionally, the Spearman Rank correlation of the CTBS scores with the PSSA scores for students below the 35th percentile was not statistically significant ($r_s = 0.22, p = 0.288$). Therefore, 2nd grade students did not statistically improve their mathematical ranking as demonstrated by the PSSA mathematics test scores in 11th grade. Figure 2 helps to demonstrate that students scoring low on the 2nd grade CTBS may have scored high or low on the 11th grade PSSA, and

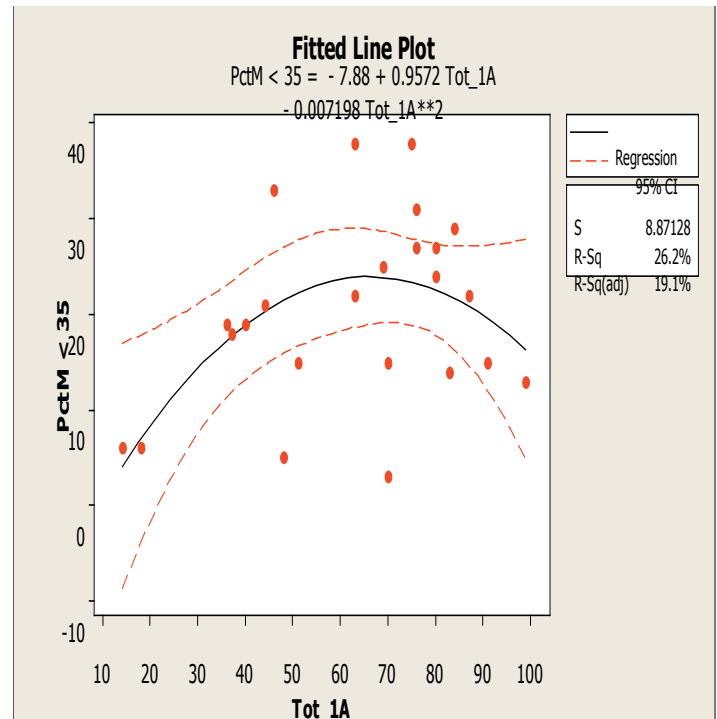


Figure 2: Scores for the low-performing students.

students scoring high on the 2nd grade CTBS may have scored high or low on the 11th grade PSSA. There was no statistically significant predictable change in mathematics performance from 2nd to 11th grade despite both instructional and curricular changes

to improve the performance of this at-risk group of students.

Discussion, Summary, and Conclusions

While it has been reported that the PSSA (Pennsylvania System of School Assessment) is a reliable and valid test (HumRRO, 2004), the data that was analyzed in the HumRRO report did not examine at-risk groups of students. However, at-risk students have been the single most important target for improving student performance (No Child Left Behind, 2004). The results of this study were that CTBS scores reported in 2nd grade and PSSA

mathematics scores reported for the same students in 11th grade were normally distributed, positively correlated, and the PSSA scores statistically predictable.

However, additional results of this study were that CTBS scores reported in 2nd grade and PSSA mathematics scores reported for the same students in 11th grade were not normally distributed and not positively correlated, and the PSSA scores were not statistically predictable for a subset of at-risk students who performed at or below the 35th percentile on the PSSA. Therefore, the results of this study question the practice of using 2nd grade CTBS scores as a driving force to change instructional methods by teachers and to make curricular changes for at-risk students in order to improve state mandated high-stakes test scores, such as the scores from the PSSA that are used in later grades.

A limitation of this study is that it was one random sample from a pool of twenty school districts in northwestern Pennsylvania. More school districts need to be sampled. More student scores need to be analyzed. Additionally, diverse stratified samples need to be made in order to determine how state mandated tests impact different potentially at-risk groups of students.

The implications of this study are two-fold. First, school districts need to be cautious about using the results of standardized mathematics test scores in the elementary grades as predictors of the results of standardized high-stakes mathematics tests in higher grades for at-risk groups of students. Secondly, curricular and instructional changes in mathematics should be made with caution for at-risk groups of students when targeting improvement of standardized high-stakes mathematics test scores in higher grades.

References

- Corbett, H. D., and Wilson, B. L. (1991), *Testing reform and rebellion*, Norwood, NJ: Ablex.
- Comprehensive Test of Basic Skills (CTBS), Monterey, CA: McGraw-Hill.
- Darling-Hammond, L., and Wise, A. E. (1985), "Beyond standardization: State standards and school improvement," *The Elementary School Journal*, 85, 315-336.
- Gilman, D. A., and Reynolds, L. A. (1991), "The side effects of statewide testing," *Contemporary Education*, 62, 272-278.
- Human Resources Research Organization (HumRRO). (2004), "PSSA issues and recommendations" (Report No. FR-04-34), Alexandria, VA: Author.
- Jones, K., and Whitford, B. L. (1997), "Kentucky's conflicting reform principles: High-stakes school accountability and student performance assessment," *Phi Delta Kappan*, 7, 276-281.
- Madaus, G. (1988a), "The distortion of teaching and testing: High-stakes testing and instruction," *Peabody Journal of Education*, 65, 29-46.
- _____ (1988b), "The influences of testing on the curriculum," in L.N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education*, pp. 83-121, Chicago: University of Chicago.
- No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
- Olson, L. (2002, April 24), "Survey shows state testing alters instructional practice," *Education Week*, p. 14.
- Popham, W. J. (1987), "The merits of measurement driven instruction," *Phi Delta Kappan*, 68, 679-682.
- Pennsylvania System of School Assessment (PSSA). Retrieved October 24, 2005, from http://www.pde.state.pa.us/a_and_t/site/default.asp?g=0&a_and_tNav=%7C630%7C&k12Nav=%7C1141%7C
- Shepard, L. A. (1989, March), "Inflated test score gains: Is it old norms or teaching the test?" *Effects of testing project*, Washington, DC: Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED334204).
- Smith, M. L. (1991), "Put to the test: The effects of external testing on teachers," *Educational Researcher*, 20, 8-11.
- Weisstein, E. W. (2005), "Anderson-Darling statistic," MathWorld--A Wolfram Web Resource. Retrieved from <http://mathworld.wolfram.com/Anderson-DarlingStatistic.html>.