Running head: HIGH STAKES TESTING AND LEARNING

High Stakes Testing and Student Learning

Bruce Thompson

Rader School of Business

Professor of Management

Milwaukee School of Engineering

(414) 277-7378

FAX: (414) 277-7479

thompson@msoe.edu

Abstract

Two recent articles by Amrein and Berliner have examined the connection–or lack thereof–between states with high-stakes testing and graduation requirements and the progress of students on other measures of achievement. These articles have generated controversy between advocates and opponents of state tests, as well as considerable interest in the popular press. The present article suggests that statistical analysis techniques are preferable to the counting approach used by Amrein and Berliner. Using the lists of high stakes states developed by Amrein and Berliner, I use both regression and analysis of variance to explore possible connections between high stakes and changes in scores on the NAEP, ACT, SAT, and Advanced Placement tests. In this analysis, there is a positive relationship between high stakes and changes in scores on the NAEP. There is a negative relationship between high stakes and changes in ACT scores. There is no evidence of relationships between high stakes and changes in scores on either the SAT or Advanced Placement test. The possible implications of these results, if confirmed by further study, are discussed.

High Stakes Testing and Student Learning:

Introduction

Articles by Amrein and Berliner (2000, March and December) have helped fuel the controversy over high-stakes testing and student achievement and received considerable attention (for examples, see Winter 2002 or Viadero 2003). These articles conclude that high stake testing either has no effect or a negative effect on student achievement as measured by national standardized tests such as the National Assessment of Educational Progress (NAEP), college entrance tests (ACT and SAT), and Advanced Placement (AP) exams.

These conclusions have been embraced by critics of high stakes and of standardized tests. Ironically they have also been welcomed by some who believe the present public educational system is beyond salvage and should be replaced by a completely market-driven system (Tucker, 2003).

These articles have been criticized by those who believe that high stakes tests are needed to drive educational improvement (Green and Forster, 2003 and Raymond and Hanushek, 2003). Much of the criticism has focused on several issues:[1]

- Bias of the sponsors and investigators. The research was sponsored by the Great Lakes Center for Educational Research and Practice whose members include National Education Association affiliates from Illinois, Indiana, Michigan, Minnesota, Ohio and Wisconsin. NEA unions are well-known critics of high-stakes testing.

It is also clear from the Amrein & Berliner articles that they are not agnostic when approaching high stakes testing. Their objections and concerns extend far beyond the areas they investigated. Berliner, in particular, seems to be part of a network of people who have long opposed the tests.

- Bias of the sample. As Amrein & Berliner themselves point out, the states giving high-stakes tests differ in many ways from other American states. In particular, these states have higher poverty and minority enrollment, two areas that research consistently shows have a negative relationship to student achievement. Thus, they hardly represent a random sample of all states.

In response, supporters of the articles have pointed out that many researchers in education have taken advocacy positions, including some of the critics. In an ideal world, perhaps, educational researchers would be completely neutral as to their conclusions. Given the realities, however, it may be more realistic to ask that the model used be valid and that opportunities for bias in the selection of data be avoided as much as possible.

In addition, to insist on randomized samples would severely limit educational research. In many cases, that may be an ideal impossible to achieve. Legislators, parents, and educators are unlikely to hold still while educators design a carefully randomized experiment to test the effects of high stakes. Insisting on an ideal study may mean that available data is ignored while waiting for

perfect conditions. An alternative approach is to encourage a variety of imperfect studies using available data, while recognizing the possibility of bias and treating the results with caution.

In the absence of high stakes, the high-stakes states' different demographic characteristics would be expected to result in differing student achievement compared to other states. However, it is not clear how these differences would affect changes in student achievement over time, which is what Amrein & Berliner attempt to measure. It is unclear whether the higher poverty in minority enrollment, for example, would make it easier or harder to raise student achievement.

Rather than either researcher or sample bias, my primary interest here is analytical methodology. Amrein & Berliner describe their methodology as "archival time series" but it is basically a counting approach. They first compare the change in scores in a particular state to the average change. They then count the number of states with scores above, below, and the same as the average change. They similarly count states with rising or falling participation rates.

In their methodology, the only consideration is whether the state exceeds or lags the average, not the magnitude of that difference. Thus, a small increase on one test could cancel a large decrease on another, or vice versa. In addition, any states with increased scores accompanied by decreased participation--or decreased scores accompanied by increased participation--were excluded, on the premise that increased participation depresses scores. Amrein & Berliner then classified the results as "weak" or "strong" depending on whether the several indicators for a state moved in the same or different directions.

This counting approach seems to have several weaknesses compared to statistical analysis:

- It ignores the magnitude of the changes. Very small changes have the same effect as very large ones. In essence, it converts continuous data into binary data resulting in a loss of information.

- There is no opportunity to generate any of the statistical values that help investigators judge the strength or weakness of the results.

- Potentially it creates a constant need for judgments about how to treat each piece of data on test scores and participation rates. This in turn raises the concern that these decisions may be either consciously or unconsciously influenced by the researchers' view of what the results should be. While considerable judgment may enter into design of a statistical analysis, the treatment of each individual datum is typically mechanical and "untouched by human hands."

Method

In this discussion I look at two alternative statistical models to examine the relationship between the implementation of high stakes and changes in scores on the NAEP, SAT, ACT, and AP tests. The first approach uses regression to analyze whether there are differences in the expected changes in the high stakes tests compared to states as a whole. The second uses analysis of variance (ANOVA) to examine the differences between average gains between the high and low stakes states and whether these differences are significant.

Test Scores

While the details of the model used varied somewhat from one test to another, depending on data availability, the first step was to calculate the change over time for each state's test scores. The following points discuss the approach taken with each test:

- National Assessment of Educational Progress (NAEP) Tests. Statewide NAEP test results are reported for individual states over several years in fourth-grade reading and mathematics and in eighth grade mathematics.[2] Not every state participated every year, so some states have missing years and others have no data. If states had at least two years of data, I calculated the average change per year in score. For states with three or more years, I calculated the average change by finding the slope of a regression line. For states with only two years of data, the change was calculated by taking the differences.

  On the fourth grade NAEP reading test, I made one additional adjustment. On average, scores declined in the second year and then returned to their former values. This fluctuation is not problematic if all states have data for every year. But depending on which year was missing, it served to either punish or reward states with missing data. To correct the missing data effect I adjusted the state results by the variation in national averages.

- SAT. I used the difference between the average 2001 and 1991 SAT I Verbal and Math scores for each state.[3]

- ACT. I used the difference between the composite ACT scores in 2001 and those in 1994 for each state.[4]

- Advanced Placement Tests. For the Advanced Placement tests, I took each state's difference between the percentage of students getting a score of three or better in 2002 and in 1997.[5]

Classifying states

I compared these changes to the lists of high-stakes states that Amrein and Berliners offer in their two articles. Apparently there is a significant element of judgment in the decision as to whether to classify a state as high stakes. In their first article (2002, March) they list eighteen states, compared to twenty-eight in their second (2002, December). In contrast, the number of stakes listed declines for many of the states in the second report compared to the first.

Analytical approach

I took two somewhat different approaches to analyzing the relationship between test score changes and states' classification as high stakes:

- In the first, I regressed the changes against an index of high-stakes, using the list from Amrein and Berliner's first article (2002, March). I tried regressing on both the number of high stakes (which varied from one to six) and a binary index of high stake states (either zero or 1). The results of the regression did not appear highly influenced by this choice, so all results included here are based on the binary index. I then examined the results to see

whether the slope was positive or negative and whether the magnitude of the slope was statistically significant at the 95% level.

- In the second approach, I grouped states' test scores into two lists, depending on whether or not they were classified as high stakes in the second article. I then applied analysis of variance (ANOVA) to measure whether the two lists were statistically different.

Results

First analysis

Table 1 shows the results of regressing the changes in states' test scores against a binary variable that had the value of one if the state was classified as having high stakes in the first Amrein & Berliner (2002, March) article and zero otherwise. I looked at two results: whether the slope of the relationship was positive or negative and the p-value of the slope, indicating whether or not the slope differed significantly from zero.

Table 1. Regression: Scores vs. High Stakes

	Slope	p-value
NAEP: 4th Grade Reading	Positive	0.16
NAEP: 4th Grade Math	Positive	0.04
NAEP: 8th Grade Math	Positive	0.14
SAT Verbal	Flat	0.6
SAT Math	Flat	0.6
ACT	Negative	0.11
Adv. Placement	Flat	0.75

As shown in Table 1, there is a positive relationship between changes in scores on the three NAEP tests and classification of these states as high stakes. In the case of the fourth grade mathematics tests this relationship is statistically significant at the 95% level.

With the exception of the ACT test, however, changes on the various tests taken by high school students seem unrelated to whether or not a state was classified as high stakes. For the ACT, the relationship seems to be negative, although not sufficiently strong to be considered statistically significant under most common standards.

Second analysis

Following publication of Amrein and Berliner's second article, I decided to repeat the analysis, using the list of high-stakes states included in that article. Following their lead, I classified states as low-stakes for the NAEP if their only high-stakes requirement was a high school graduation test. For the various high school tests (SAT, ACT, and AP), a state was classified as high stakes only if it had a high-school graduation requirement. Several states with newly-introduced high-stakes requirements were classified as low stakes if in the judgment of Amrein & Berliner (2002, December), the stakes were introduced too recently to have had an effect.

Another difference between the two analyses lies in the treatment of ACT and SAT tests. In Table 1, the results for these tests were calculated using all states. But most states are dominated by one or the other of these two tests. In the second analysis, I used ACT results only from ACT-dominant states and SAT results only from SAT-dominant states.

Table 2 shows the number of states included in each of the analyses.

Table 2. Number of States		
	Low Stakes	**High Stakes**
NAEP: 4th Grade Reading	21	21
NAEP: 4th Grade Math	22	23
NAEP: 8th Grade Math	23	21
SAT Verbal	15	10
SAT Math	15	10
ACT	19	7
Adv. Placement	33	18

Despite the differences in both the analytical technique and the number of states classified as high stakes, the results from Tables 1 and 3 are remarkably consistent. In Table 3 high-stakes states do better on the three NAEP tests (significantly better on the two mathematics tests), worse on the ACT, and are in a statistical dead heat on the others.

Table 2. Analysis of Variance		
	Higher	**p-value**
NAEP: 4th Grade Reading	High stakes	0.11
NAEP: 4th Grade Math	High stakes	0.015
NAEP: 8th Grade Math	High stakes	0.13
SAT Verbal	Tie	0.6
SAT Math	Tie	0.9
ACT	Low stakes	0.11
Adv. Placement	High stakes	0.26

Participation effects

An issue related to test scores is participation rates. It is widely reported, for example, that some schools encourage weaker students to not take state-required tests, or find reasons to exempt students expected to do poorly. Likewise, at the high school level it is assumed that states with

low SAT or ACT participation, for example, have an advantage over high participation states since only students aiming at highly competitive out-of-state colleges will take the test.

In their tabulations of states with increased or decreased test scores, Amrein and Berliner eliminate any increase in a score accompanied by a decrease in participation, or decrease in score along with an increase in participation. This rule has the effect of eliminating a majority of the results from their count and a nontrivial effect on the outcome. For example, in their second report, I count that 23 of their results favor high-stakes states while 32 go against these states. If participation is not considered, the tabulation rises to 63 favoring high-stakes states and 47 against them.

To get an idea of the possible effect of participation rates on test scores, I examined participation rates on two of the tests: the percentage change in graduates tested on the ACT and changes in the number of advanced placement exams per thousand eleventh and twelfth graders. In the ACT there was a slight negative relationship between increases in participation and decreases in test scores, but it was far from statistically significant (p-value around .8). With the AP exams there was no hint of a relationship.

The implicit model behind throwing out gains accompanied by reduced participation is that weaker students determine participation rates. These results suggest that this model needs further testing before it is accepted as applying to all tests. Advanced Placement enrollment, for example, may have expanded in many states primarily through the involvement of more schools offering more courses, rather than by enrolling more marginal students. All of these tests (particularly the

NAEP) are low stakes for the schools, so there is little incentive to limit participation among students expected to score poorly.

Test timing

Another issue mentioned both by Amrein & Berliner and by some of their critics is that of timing. The tests must be carefully chosen, it is argued, to bracket the imposition of high stakes. Ideally the first test used would have been given before the start of high stakes and the last after they have fully taken effect. Trying to properly time the measurements, however, is difficult. First, the number of times the NAEP, in particular, has been offered at the state level is limited, making proper before and after timing difficult. Second, in many states the high stakes were introduced over a period of years. Third, trying to match particular test dates for each state makes any sort of statistical analysis very difficult. Fourth, judging when the requirement took effect can be tricky: is it when first proposed, when passed by the legislature, on the legal effective date, when teachers made changes to accommodate the change, or when all students taking a test would have lived through the new requirements? Finally, trying to tailor the results to effective dates seems to be an open invitation for investigators to micromanage the data, adjusting data so they fit a model.

I would argue that the timing issue is largely irrelevant. If high stakes, or any changes, do impact student learning, that effect should appear over a period of time as teachers adjust their teaching and students spend more time under the new regime.

Discussion

It is worth reiterating that the high-stakes states do not represent a random sample of all states. They differ from the average in many ways, ranging from those known to impact student outcomes, notably poverty and minority enrollment, to others whose impact is purely speculative, such as the degree of centralization of state power. My results are suggestive but hardly conclusive. The differences may result from some factor completely unrelated to the imposition of high stakes.

Conclusions

Given that caveat, here is what the results suggest:

1. High stakes do seem to have a positive impact on average elementary and middle school students, particularly in mathematics, as shown by the fourth and eighth grade NAEP tests. Thus these results support the hopes of those who believe that high stakes will improve the learning of ordinary students in basic subjects.

2. The results for the tests given at high school are more ambiguous. In all but one case, there seems to be little or no relationship. This makes some sense because the high stakes tests are minimal competency tests. Students in an advanced placement class or taking the SAT, and their teachers, probably have little worry about meeting a state's requirements. The exception, of course is the ACT, where there is some evidence of a possible negative effect.

A theory of high stakes effects

One theory about the results of high-stakes testing, suggested both by some supporters and by some opponents, is that they raise performance among lower-performing students while depressing performance at the higher end. In this theory, high stakes can act as both a floor and ceiling on student achievement. My analysis gives support to the first part of this theory, while supporting the second part only in the case of the ACT.

In this theory, high stakes encourage two changes in teachers' behavior that result in increased performance among lower-performing students. First, it encourages them to concentrate on the basic reading and mathematics skills that appear on the test, and in which these students are deficient. Second, particularly if schools are judged based on the percentage of students rated proficient, it encourages teachers to concentrate their efforts on the students who are unlikely to reach proficiency without additional help.[6]

This theory explains the ceiling effect as the mirror image of the floor effect: teachers ignore both more advanced material not given on the test and more advanced students who are expected to score well no matter what.

Why would the ceiling effect show up with the ACT and not the SAT? A possible explanation may be different test designs. The SAT I is described as a test of critical reading and problem solving, a modified aptitude test designed in part to identify "diamonds in the rough," students from substandard high schools with the talent to succeed in college. The ACT was started in 1959 as a curriculum-based achievement test (Perry, 2002) and as a reaction to the SAT. To the extent

it measures aptitude rather than material in the curriculum, the SAT may be less vulnerable to what goes on in high school.

Likewise, state high stakes may be largely irrelevant to Advanced Placement tests. The typical AP class is likely to consist of students who have either satisfied or are assured of meeting any state requirements. The high stakes facing them are the AP tests themselves. AP teachers need not concern themselves with their students meeting the state requirements.

Thus, these results are consistent with a floor-ceiling model of the effect of state high stakes. If a state finds its high stakes are improving performance at the bottom but damping the top, the model suggests several steps it can take to remove the ceiling effect while keeping the advantages of a rising floor.

A possible strategy

If the low level of the test is the problem, a possible solution would be to give students differing versions of a test based on their abilities. Computerized adaptive testing, in which question difficulty is adjusted depending on a student's success in answering the first few questions is one means to do this. With written tests, it is possible to give students in a grade different versions of the test depending on their performance either on a pretest or on the previous regular test administration. Either approach, of course, increases the administrative burden of testing.

If a cause is teachers concentrating on struggling students to the detriment of those ready for more challenging material, offering a spectrum of programs tailored to individual student needs

may help. Thus where the needs of students vary widely, a strategy of grouping students according to their needs and designing programs that fit those needs may better avoid the floor-ceiling effect, so that students in danger of not meeting state standards will get the help they need, while students already meeting the standards will be free to tackle more challenging material.

The increasing popularity of advanced placement and International Baccalaureate programs at the high school level seems a tacit recognition of the need for programs that challenge students who have no difficulty meeting the minimal standards in high stakes. In a recent study of longitudinal test scores in a large urban school system (Thompson, 2003), I found that the average student's mathematics and reading scores actually declined slightly in ninth grade. Since ninth grade is a transition year, the average high school may know little about the incoming freshmen, assigning students to classes where teachers struggle to bring up the students who are seriously behind while neglecting the needs of the majority ready for more advanced material. By tenth grade, when scores start moving up again, students are more likely to be assigned to classes appropriate to their needs.

There is a growing wealth of data on educational outcomes. With the widespread availability of statistical tools, it seems desirable to use these tools to tease out the stories these data can tell. No one analysis is likely to be definitive, but gradually a better picture will emerge as to what works in education and what does not.

References

Amrein, A.L. & Berliner, D.C. (2002, March 28). High-stakes testing, uncertainty, and student learning Education Policy Analysis Archives, *10*(18). Retrieved 14 January 2003 from http://epaa.asu.edu/epaa/v10n18/

Amrein, A.L. & Berliner, D.C. (2002, December). The Impact of High-Stakes Tests on Student Academic Performance, Tempe, AZ : Arizona State University Education Policy Research Unit (EPRU). Retrieved 14 January 2003 from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf

Becker, W.E., & Rosen, S. (1992). The learning effect of assessment and Evaluation in high school, Economics of Education Review, 11(2), 107-118.

Greene, J. P. , & Greg Forster, G. (2003, January). Burning High-Stakes Testing at the Stake, in the Education Gadfly. Dayton OH: Thomas B. Fordham Foundation. Retrieved 14 January 2003 from http://www.edexcellence.net/gadfly/v03/gadfly01.html#jaygreg1.

Perry, D. A., Chair (2002). The Use of Admissions Tests by the University of California Retrieved February 14, 2003 from http://www.ucop.edu/news/sat/boars.pdf.

Raymond, M.E., and Hanushek, E.A. (2003), High stakes research, Education Now/Summer 2003, 48-55.

Thompson, B. R. (In press). What Do Students Learn in High School? Using Student Gain Scores to Evaluate Student Progress, Milwaukee: Milwaukee School of Engineering.

Tucker J. (2003, January). Another Central Plan Fails, Ludwig von Mises Institute. Retrieved 14 January 2003 from http://www.mises.org/fullstory.asp?control=1130 .

Viadero, D. (2003, January 8) Reports Find Fault With High-Stakes Testing. Education Week.

Winter, Greg (2002, December 28). More Schools Rely on Tests, but Study Raises Doubts, New York Times. Retrieved 14 January 2003 from http://www.nytimes.com/2002/12/28/education/28EXAM.html.

.

Endnotes

[1] Raymond and Hanushek (2003) also propose a statistical model of NAEP changes in high stake states versus other states, concentrating on the growth of scores between the fourth grade tests in 1996 and the eighth grade test in 2000. The find a significant advantage for the high stakes states in mathematics. They did not seem to explore the high school tests included by Amrein and Berliner.

[2] National Assessment of Educational Progress Data is available at http://nces.ed.gov/nationsreportcard/. While tenth grade mathematics and reading tests have been given nationally for many years, there are no state by state breakdowns. Only one year of state data has been published for eighth grade reading, allowing no comparisons over time.

[3] SAT scores are available at www.collegeboard.com.

[4] ACT scores are available at http://www.act.org/.

[5] AP data are available at www.collegeboard.com.

[6] As described, this theory ignores the possible effect of high-stakes tests on student motivation. See Becker & Rosen (1992) for a discussion of the effect of assessment on motivation.