

Copyright is retained by the first or sole author, who grants right of first publication to the **EDUCATION POLICY ANALYSIS ARCHIVES**. EPAA is a project of the [Education Policy Studies Laboratory](#).

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the [ERIC Clearinghouse on Assessment and Evaluation](#) and are permanently archived in *Resources in Education*.

Volume 11 Number 24

August 4, 2003

ISSN 1068-2341

High-Stakes Testing: Another Analysis **Barak Rosenshine** **University of Illinois at Urbana, Champaign**

Citation: Rosenshine, B. (2003, August 4). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, 11(24). Retrieved [date] from <http://epaa.asu.edu/epaa/v11n24/>.

Abstract

Amrein and Berliner (2002b) compared National Assessment of Educational Progress (NAEP) results in high-stakes states against the national average for NAEP scores. They studied NAEP scores for 8th grade mathematics, 4th grade mathematics, and 4th grade reading. They concluded that states that introduced consequences (high-stakes) to their statewide tests did not show any particular gains in their statewide NAEP scores. However, there was no comparison group in their analysis. In this analysis, a comparison group was formed from states that did not attach consequences to their state-wide tests. This analysis showed that states that attached consequences outperformed the comparison group of states on each of the three NAEP tests for the last four-year period. These results showed that, overall, these was a meaningful carryover from attaching consequences on statewide tests to statewide NAEP scores.

As reported by Viadero (2003a, 2003b), Audrey Amrein and David Berliner (2002a, 2002b, 2002c) have studied the effects of attaching consequences or accountability ("high stakes") to student scores on their statewide exams. These consequences include monetary awards to schools or teachers, authority to replace a principal or

a teacher, and limiting grade-to-grade promotion. Individual states imposed one to six such consequences, with an average of three consequences. (See Table 1 in Amrein and Berliner (2002b).)

In an admirable decision, Amrein and Berliner did not look at each state's scores on their own statewide tests. Rather, they looked at each state's scores on an independent measure, the National Assessment of Educational Progress. They compared the four-year changes in NAEP scores in each high-stakes state against the average changes for all the states that took each NAEP test. They studied NAEP results in three areas: 8th grade mathematics, 4th grade mathematics, and 4th grade reading.

They concluded that, there were "no consistent effects across states" after consequences were introduced. (Amrein & Berliner, 2002b, p. 57). Some states had larger increases than the national average, but the NAEP changes in other states were less than the national average. They concluded that students in high-stakes states were not learning anything beyond the specific content of the statewide tests.

Their analysis, however, did not include a comparison group. The economist John H. Bishop, quoted in Viadero (2003b), said "The natural thing to do would be to compare the states that had accountability systems to ones that didn't." This reanalysis follows Bishop's suggestion and is a comparison of the NAEP gains in the high-stakes states against the NAEP gains in states that did not have statewide accountability procedures.

Not all the 26 high-stakes states were included in the Amrein and Berliner analysis. They noted that some states may have manipulated their NAEP scores by exempting some of the special education students and students with limited English proficiency from those taking the NAEP test. They excluded these states or individual results from these states. For example, all the results from North Carolina and Texas, two states with large NAEP increases, were labeled as "unclear" and were not included in their analyses. All of their reported results are for the remaining 8 to 12 "clear" states (depending upon the specific NAEP exam). I believe this separation into "clear" and "unclear" high stakes states was a valuable step because it allows the analysis to focus only on the results in the clear states.

My analysis uses the NAEP four-year gain from these same 8-12 "clear" high-stakes states. An additional 14 to 18 states (depending upon the specific NAEP exam) did not attach consequences to tests and they served as the comparison group in my analysis. My analysis is based on the gain, from cohort to cohort, between 1996 and 2000 for the two mathematics tests and on the gain between 1994 and 1998 on the reading test. One could, of course, go back eight years, but then the number of high-stakes states would be much smaller. The tables from which these numbers came are readily available at the NAEP website (<http://nces.ed.gov/nationsreportcard/>).

The results in the Table 1 show the average NAEP increases in the "clear" high-stakes states were much higher than the increases in the comparison states. . In 8th grade mathematics and in 4th grade reading the mean increase for the clear high-stakes states was double the increase for the states without consequences.

The effect sizes for the comparisons were .35 for 4th grade math, .79 for 8th grade math, and .61 for 4th grade reading. These effect sizes have been called moderate to large. An effect size of .35 means that the average high-stakes state would score at the 63rd percentile of the comparison states. Effect sizes of .79 and .62 correspond to the 78th and the 73rd percentile of the comparison states.

Table 1
Average four-year increases on NAEP scores for
Clear High-Stakes States and for States without High Stakes

NAEP Test	Average four-year increase in clear high-stakes states	Average four-year increase in states without high-stakes
4th grade mathematics (1996-2000)	3.45 (n=11 states)	2.40 (n=15 states)
8th grade mathematics (1996-2000)	3.42 (n=7 states)	1.63 (n=13 states)
4th grade reading (1994-1998)	3.44 (n=9 states)	1.21 (n=14 states)

These results suggest that students in the clear high-stakes states were, indeed, learning mathematics and reading that was beyond the specific content of the statewide tests. Their NAEP achievement, on all three tests, was higher than the achievement of students in NAEP states that did not attach high-stakes to their statewide exams.

These results might lead us to reconsider some of the textual statements in the Amrein and Berliner report (2002b). Amrein and Berliner wrote that "...the imposition of high-stakes testing results in a more narrow form of training..." (p. 6) Perhaps, and perhaps not. But this apparent narrow training did not prevent students in the clear high-stakes states from doing quite well on the NAEP tests, better than the students in states that did not attach consequences to statewide tests.

Did state NAEP scores really decrease?

The Amrein and Berliner (2002b) report states that NAEP scores in many high-stakes states "decreased" after consequences were implemented. They wrote that "grade 4 reading achievement decreased" (p.19) in Alabama after stakes were attached to statewide tests. They wrote that in Nevada, "grade 4 math achievement decreased." But "decrease" is a relative term. Scores in Nevada increased three points on the 4th grade NAEP math test between 1996 and 2000. But because this increase was less than the national average of four points, Nevada was listed as "grade 4 math achievement decreased." Grade 4 reading achievement increased by three points in Alabama between 1994 and 1998, the same increase as the

national average. However, Amrein and Berliner noted that the percentage of students exempted from NAEP increased, in Alabama, between 1990 and 1994. Therefore, they concluded that "after stakes were attached to tests in Alabama, grade four reading achievement decreased." (p. 19)

Overall, Amrein and Berliner reported that NAEP scores in the 4th grade mathematics test "decreased" in 8 states. However, there was no actual decrease in any of these states: one state had no change between 1996 and 2000 but the remaining 7 states actually showed increases of from 1 to 4 points during this period. If one adds them up, across the three NAEP exams, there was an absolute decrease in only two of the high-stakes state scores used in the Amrein and Berliner analysis and an absolute decrease in 10 of the NAEP scores for states that did not include accountability measures.

Discussion

This analysis showed that the clear high-stakes states outperformed the comparison states on each of the three NAEP tests for the last four-year period. If my analytic approach makes sense, and if these results are confirmed by others, then I hope we can begin to study what these results mean

The results showed there was a meaningful carryover effect, in some states, from statewide testing to the NAEP. Based on these results, it is not appropriate for Amrein and Berliner (2002b) to say that attaching accountability to statewide tests "results in a narrow form of training," or "high-stakes testing creates a 'training effect' only." (p. 6) Nor is it appropriate to say that "students were learning the content of the state-administered test and perhaps little else." (p. 60)

Although attaching accountability to statewide tests worked well in some high-stakes states, it was not an effective policy in all states. South Carolina, Massachusetts, and Alabama did particularly well in 4th grade math, but not New Mexico, West Virginia, or Kentucky. Indiana and Alabama did particularly well in 8th grade math, but not New Mexico or Missouri. Louisiana, Delaware and Virginia did particularly well in 4th grade reading, but not Missouri or New Mexico.

It would be appropriate to study these successful and less successful high-stakes states and learn how they achieved their results. It would be less appropriate to simply use these results as a hammer and blindly require all states to impose consequences.

I find it unlikely that the NAEP results in the high-scoring states were obtained merely because two weeks were devoted to test preparation. I find it unlikely that the NAEP results in the high-scoring states were obtained only because consequences and accountability were introduced.

My guess would be that there is a strong academic focus in these classrooms and these schools. The research has supported academically-focused classrooms since 1960, and having seen lots of trivia in classrooms, I welcome a return to an academic focus. I've been in many low-income high-achievement elementary schools and they are, indeed, high achieving places. I have seen history projects, discussion of novels, Junior Great Books, impressive mathematics lessons, and

tutoring during lunch hours. It would be unfair to the efforts of students, teachers, and principals to say that they are merely focusing on statewide tests.

Statewide policies and district policies may also facilitated these increases. If so, how was that accomplished? Did these policies affect all schools? What forms of help did the state and the district provide for the classrooms? All these items may be worthy of study. We might also reexamine the lists of statewide consequences in Table 1 of Amrein and Berliner (2002b) and ask which of these consequences might act as a motivating and not a threatening factor.

Audrey Amrein and David Berliner have performed an important service by focusing on the consequences that some states have attached to statewide testing. I think their use of state NAEP scores as an independent assessment was a brilliant move. Their use of NAEP scores also allows others to conduct additional analyses of this public data. My additional analysis suggests that students in some high-stakes states have done very well on the NAEP tests. I hope we can now study how this happened.

Acknowledgement

I was in contact with Audrey Amrein throughout this reanalysis and it could not have been done without her help. She provided clear and full answers to each of my frequent email questions. . Some of my ideas for this analysis came as a result of email discussions with David Berliner, and I thank him. Sam Stringfield, Burnie Bond, Bob Stevens, Jere Brophy and Marilyn Kohl all provided useful questions and comments.

References

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

Amrein, A.L. & Berliner, D.C. (2002b). The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP Test results in states with high school graduation exams . Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved July 18, 2003 from <http://www.asu.edu/educ/eps/EPRU/documents/EPSSL-0211-126-EPRU.pdf>.

Amrein, A.L. & Berliner, D.C. (2002c). An analysis of some unintended and negative consequences of high-stakes testing. Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved July 18, 2003 from <http://www.asu.edu/educ/eps/EPRU/documents/EPSSL-0211-125-EPRU.pdf>.

Viadero, D. (2003a, January 8). Reports find fault with high-stakes testing. *Education Week*, 22(16), 5.

Viadero, D. (2003b, February 5). Researchers debate impact of tests. *Education Week*, 22(21), 1 & 12.

About the Author

Barak Rosenshine

College of Education
University of Illinois, Urbana—Champaign

Email: rosenshine@uiuc.edu

Barak Rosenshine is an emeritus professor of educational psychology at the University of Illinois at Urbana-Champaign. His research speciality is classroom instruction.

The World Wide Web address for the *Education Policy Analysis Archives* is
epaa.asu.edu

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, [Gene V Glass, glass@asu.edu](mailto:glass@asu.edu) or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

[Michael W. Apple](#)
University of Wisconsin

[Greg Camilli](#)
Rutgers University

[Sherman Dorn](#)
University of South Florida

[Gustavo E. Fischman](#)
California State University—Los Angeles

[Thomas F. Green](#)
Syracuse University

[Craig B. Howley](#)
Appalachia Educational Laboratory

[Patricia Fey Jarvis](#)
Seattle, Washington

[Benjamin Levin](#)
University of Manitoba

[David C. Berliner](#)
Arizona State University

[Linda Darling-Hammond](#)
Stanford University

[Mark E. Fetler](#)
California Commission on Teacher Credentialing

[Richard Garlikov](#)
Birmingham, Alabama

[Aimee Howley](#)
Ohio University

[William Hunter](#)
University of Ontario Institute of Technology

[Daniel Kallós](#)
Umeå University

[Thomas Mauhs-Pugh](#)
Green Mountain College

[Les McLean](#)
University of Toronto

[Michele Moses](#)
Arizona State University

[Anthony G. Rud Jr.](#)
Purdue University

[Michael Scriven](#)
University of Auckland

[Robert E. Stake](#)
University of Illinois—UC

[Terrence G. Wiley](#)
Arizona State University

[Heinrich Mintrop](#)
University of California, Los Angeles

[Gary Orfield](#)
Harvard University

[Jay Paredes Scribner](#)
University of Missouri

[Lorrie A. Shepard](#)
University of Colorado, Boulder

[Kevin Welner](#)
University of Colorado, Boulder

[John Willinsky](#)
University of British Columbia

EPAA Spanish Language Editorial Board

Associate Editor for Spanish Language
[Roberto Rodríguez Gómez](#)
[Universidad Nacional Autónoma de México](#)

roberto@servidor.unam.mx

[Adrián Acosta \(México\)](#)
Universidad de Guadalajara
adrianacosta@compuserve.com

[Teresa Bracho \(México\)](#)
Centro de Investigación y Docencia
Económica-CIDE
bracho dis1.cide.mx

[Ursula Casanova \(U.S.A.\)](#)
Arizona State University
casanova@asu.edu

[Erwin Epstein \(U.S.A.\)](#)
Loyola University of Chicago
Epstein@luc.edu

[Rollin Kent \(México\)](#)
Universidad Autónoma de Puebla
rkent@puebla.megared.net.mx

[Javier Mendoza Rojas \(México\)](#)
Universidad Nacional Autónoma de
México
javiermr@servidor.unam.mx

[Humberto Muñoz García \(México\)](#)
Universidad Nacional Autónoma de
México
humberto@servidor.unam.mx

[Daniel Schugurensky](#)(Argentina-Canadá)
OISE/UT, Canada

[J. Félix Angulo Rasco \(Spain\)](#)
Universidad de Cádiz
felix.angulo@uca.es

[Alejandro Canales \(México\)](#)
Universidad Nacional Autónoma de
México
canalesa@servidor.unam.mx

[José Contreras Domingo](#)
Universitat de Barcelona
Jose.Contreras@doe.d5.ub.es

[Josué González \(U.S.A.\)](#)
Arizona State University
josue@asu.edu

[María Beatriz Luce](#)(Brazil)
Universidad Federal de Rio Grande do
Sul-UFRGS
lucemb@orion.ufrgs.br

[Marcela Mollis \(Argentina\)](#)
Universidad de Buenos Aires
mmollis@filo.uba.ar

[Angel Ignacio Pérez Gómez \(Spain\)](#)
Universidad de Málaga
aiperez@uma.es

[Simon Schwartzman](#) (Brazil)
American Institutes for
Resesarch—Brazil (AIRBrasil)

dschugurensky@oise.utoronto.ca

[Jurjo Torres Santomé](#) (Spain)
Universidad de A Coruña
jurjo@udc.es

simon@sman.com.br

[Carlos Alberto Torres](#) (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu

EPAA is published by the Education Policy Studies
Laboratory, Arizona State University