EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Volume 11 Number 25 August 4, 2003 ISSN 1068-2341

Re-analysis of NAEP Math and Reading Scores in States with and without High-stakes Tests: Response to Rosenshine

Audrey Amrein-Beardsley
David C. Berliner
Arizona State University

Citation: Amrein-Beardsley, A. A. & Berliner, D. C. (2003, August 4). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives, 11*(25). Retrieved [Date] from http://epaa.asu.edu/epaa/v11n25/.

Abstract
Here we address the criticism of our NAEP analyses by Rosenshine (2003). On the basis of his thoughtful critique we redid some of the analyses on which he focused. Our findings contradict his. This is no fault of his, the reasons for which are explained in this paper. Our findings do support our position that high-stakes tests do not do much to improve academic achievement. The extent to which states with high-stakes tests outperform states without high-stakes tests is, at best, indeterminable. Using 1994-1998 NAEP reading and 1996-2000 NAEP math data and accounting for NAEP exemption rates for the same years, we found that states with high-stakes tests are not outperforming states without high-stakes tests in reading in the 4th grade or math in the 8th grade at a statistically significant level. States with high-stakes tests are, however, outperforming states without high-stakes tests in math in the 4th grade at a

statistically significant level. Our findings also support our earlier stance that states with high-stakes tests are exempting more students from participating in the NAEP than are states without high-stakes tests. This is more prevalent the more recent the NAEP test administration. This is illustrated in the tables below.

Introduction

In our research, we were concerned that scores on high-stakes state tests could easily be manipulated through narrowing of the curriculum, drilling on items similar to the test, increasing exclusion rates of students, increasing in dropouts and push-outs, and the like. To judge whether that concern was valid, we looked at audit tests—tests that might have some overlap with a state's own test but where school personnel were under much less intense pressure to achieve higher scores. We chose a series of audit tests to examine—SAT, ACT and AP tests, as well as all administrations of the NAEP reading and mathematics tests. We also studied whether some unanticipated side effects were present when high-stakes tests were introduced, such as increased GED taking, increased reporting of cheating, problems of teacher morale, problems with student motivation to learn, and so forth.

Substantive criticism of our work, thus far, has been limited to the NAEP data we reported. To our knowledge, the other conclusions we reached have not yet been subject to the same kinds of thoughtful criticism. So for now, given the methods that we used for analyses, our findings in those other areas stand. We concluded that there was no systematic pattern of gains on SATs, ACTs or AP exams. That is, we found no evidence of transfer from the state tests to these other tests, tests that can be considered as audit measures. In addition, we found increased drop-out rates and decreased high school graduation rates, increased rates by which students participated in the GED program, and a host of troubling negative affects associated with high-stakes testing.

Here we address the criticism of our NAEP analyses by Rosenshine (2003). On the basis of his thoughtful criticism, we redid some of the analyses on which he focused and now have a different view of the findings. What we found contradicts what we found in both of our earlier papers (Amrein & Berliner, 2002a; Amrein & Berliner 2002b) but the data analyzed below are for different years of data from those used in the earlier papers. Following the form of the analyses done by Rosenshine the data analyzed below are only for the years 1994-1998 for the NAEP reading test, and 1996-2000 for the NAEP mathematics tests. In addition, in our earlier work we used the national trend line as the contrast or control group for our analyses. In this analysis, we use the composite score for states without high-stakes tests as the control. In addition, our findings contradict the findings reported by Rosenshine. This is no fault of his. Rosenshine used our designation of clear and unclear states with and without high-stakes tests from the second of our two papers.[2] We communicated many times and approved the states he used in his analysis. Given more consideration, however, we noticed the distinctions we made between clear and unclear states was based on our overall findings which were based on all of the available NAEP data. In other words, Rosenshine analyzed the latest two NAEP administrations in reading and math using the distinctions we made between clear and unclear states when we used all of the available NAEP data, approximately 10 years of NAEP data per subject. To complicate things more, because we used the national trend line as our control group, our clear/unclear distinctions were also made factoring in the national average. Rosenshine did not do this which makes for differences in the findings. He used the states without high-stakes tests as the control. This makes for a better analysis and we have followed his lead here. In short, Rosenshine should not be faulted for his findings nor should he be considered wrong in what he did. He did a fine reanalysis of our NAEP examination given the information he had at that point, and here we are redoing his.

NAEP Reading Grade 4 1994-1998

Taking Table 1 from Amrein & Berliner (2002b) and the states in which high stakes tests were implemented before 1994 and between 1994 and 1998, we re-ran our analyses, as Rosenshine did, using all states with high-stakes tests and, as the control group, all states without high-stakes tests for which NAEP data were available. What we found in regards to reading grade 4 achievement from 1994-1998 is as follows:

Table 1

Fourth grade 1994-1998 NAEP reading scores (raw data).

States without high-stakes tests:	NAEP 1994	NAEP 1998	States with high-stakes tests:	NAEP 1994	NAEP 1998
Arizona	206	207	Alabama	208	211
Arkansas	209	209	Kentucky	212	218
California	197	202	Louisiana	197	204
Colorado	213	222	Maryland	210	215
Connecticut	222	232	Michigan	n/a	217
Delaware	206	212	Mississippi	202	204
Florida	205	207	Missouri	217	216
Georgia	207	210	New Mexico	205	206
Hawaii	201	200	North Carolina	214	217
Iowa	223	223	Oklahoma	n/a	220

Kansas	n/a	222		South Carolina	203	210	
Maine	228	225		Tennessee	213	212	
Massachusetts	223	225		Texas	212	217	
Minnesota	218	222		West Virginia	213	216	
Montana	222	226					
Nevada	n/a	208					
New Hampshire	223	226					
New York	212	216					
Oregon	n/a	214					
Rhode Island	220	218					
Utah	217	215					
Virginia	213	218					
Washington	213	217					
Wisconsin	224	224	Change in Score				Change in Score
Wyoming	221	219					
OVERALL	214.7	216.8	+2.1*	OVERALL	208.8	213.1	+4.3*

*Significant at a $p < .05$ level

Table 1 illustrates that the states with high-stakes tests outperformed those states without high-stakes tests on the NAEP grade 4 reading tests over the period 1994-1998. However, as shown in our earlier research (Amrein & Berliner, 2002a; Amrein & Berliner, 2002b), the rates by which students are excluded from the NAEP must be taken into consideration to determine whether gains and losses are clear (interpretable) or unclear (not interpretable).

Clear gains can be determined if a state's scores increase while the rates by which students are exempted from the NAEP stay the same or decrease. In other words, when the pool of students sampled to participate in the NAEP is less selective then the likelihood that their scores would increase artificially is nullified. Under these conditions such gains are clear. Clear losses can be determined if a state's scores decrease at the same time the rates by which students are exempted from the NAEP increase. In this case, the pool of students sampled was more selective and yet the scores still went down. Under these conditions it is reasonable to interpret those findings as a clear loss.

Unclear gains are the case when a state's scores increase while the rates by which students are exempted from the NAEP increase. In other words, the pool of students sampled to participate in the NAEP is more selective and therefore likely to have biased the resulting gains. If lower-scoring students are pulled from the NAEP sample, scores on the NAEP will increase. This makes for unclear results. Unclear losses are the case when a state's scores decrease at the same time the rates by which students are exempted from the NAEP sample decrease. In this case, the pool of students sampled was less selective so it is difficult to determine whether the addition of more lower-scoring students or an actual decrease in achievement caused the resulting losses.

We believe that it is absolutely necessary to make these kinds of judgments about each state because states with high-stakes tests are those states that increasingly are exempting more students from participating in the NAEP. "In states with high-stakes tests, between 0%–49% of the gains in NAEP scores can be explained by increases in rates of exclusion." (Amrein & Berliner, 2002a)

Looking simply at those states for which clear gains or losses are applicable, an analysis of the data yields the results given in Table 2. In this table states shaded in green are those for which clear results were evident, states shaded in red are those for which unclear results were illustrated, and states shaded in yellow are those for which there were not enough data to analyze gains or losses appropriately.

As can be seen, only two states included in the states with high-stakes column can be counted as states with "clear" effects. The composite data are not significant but the table illustrates the extent to which states with high-stakes tests are not gaining in score simply because of their high-stakes testing policies.

Table 2

Fourth grade 1994-1998 NAEP reading scores with states coded as clear or unclear in their gains and losses.

States without high-stakes tests:	NAEP 1994	NAEP 1998	States with high-stakes tests:	NAEP 1994	NAEP 1998
Arizona	206	207	Alabama	208	211
Arkansas	209	209	Kentucky	212	218
California	197	202	Louisiana	197	204
Colorado	213	222	Maryland	210	215
Connecticut	222	232	Michigan	n/a	217
Delaware	206	212	Mississippi	202	204
Florida	205	207	Missouri	217	216
Georgia	207	210	New Mexico	205	206

Hawaii	201	200		North Carolina	214	217	
Iowa	223	223		Oklahoma	n/a	220	
Kansas	n/a	222		South Carolina	203	210	
Maine	228	225		Tennessee	213	212	
Massachusetts	223	225		Texas	212	217	
Minnesota	218	222		West Virginia	213	216	
Montana	222	226					
Nevada	n/a	208					
New Hampshire	223	226					
New York	212	216					
Oregon	n/a	214					
Rhode Island	220	218					
Utah	217	215					
Virginia	213	218					
Washington	213	217					
Wisconsin	224	224	Change in Score				Change in Score
Wyoming	221	219					
OVERALL	215.4	217.0	+1.6*	OVERALL	209.5	210.0	+0.5

*Significant at a p < .05 level

The composite data are important in that they nullify what one might conclude looking simply at Table 1. States with high-stakes tests are not outperforming states without high-stakes tests in reading grade 4 performance. Rather, as illustrated in Table 2, states without high-stakes tests gained in reading grade 4 performance at a statistically significant level. Given only two states are included as clear states with high-stakes tests, states with high-stakes tests made insignificant gains and the differences between the two mean gains are not statistically significant.

Most importantly, what can be drawn from Table 2 is that states with high-stakes tests are exempting more students from participating in the reading grade 4 NAEP. Ninety percent of the states with “unclear” gains are states with increases in the rates by which students were exempted from the test. This supports the notion that states with high-stakes tests are not gaining in NAEP scores simply because of their high-stakes testing policies.

NAEP Math Grade 4 1996-2000

Taking Table 1 from Amrein & Berliner (2002b) and the states in which high stakes tests were implemented before 1996 and between 1996 and 2000, we re-ran our analyses using all states with high-stakes tests and, as the control group, all states without high-stakes tests for which NAEP data were available. What we found in regards to math grade 4 achievement from 1996-2000 is as follows:

Table 3

Fourth grade 1996-2000 NAEP mathematics scores (raw data)

States without high-stakes tests:	NAEP 1996	NAEP 2000	States with high-stakes tests:	NAEP 1996	NAEP 2000
Alaska	224	n/a	Alabama	212	218
Arizona	218	219	California	209	214
Arkansas	216	217	Delaware	215	n/a
Colorado	226	n/a	Florida	216	n/a
Connecticut	232	234	Indiana	229	234
Georgia	216	220	Kentucky	220	221
Hawaii	215	216	Louisiana	209	218
Idaho	n/a	227	Maryland	220	222
Illinois	n/a	225	Massachusetts	229	235
Iowa	229	233	Michigan	226	231
Kansas	n/a	232	Mississippi	208	211
Maine	233	231	Missouri	225	229
Minnesota	232	235	Nevada	217	220
Montana	228	230	New Jersey	227	n/a
Nebraska	228	226	New Mexico	214	214
North Dakota	231	231	New York	223	227
Oregon	224	227	North Carolina	224	232

Rhode Island	221	225		Ohio	n/a	231	
Tennessee	219	220		Oklahoma	n/a	225	
Utah	226	227		Pennsylvania	226	n/a	
Vermont	225	232		South Carolina	213	220	
Washington	225	n/a		Texas	229	233	
Wisconsin	231	n/a	Change in Score	Virginia	222	230	Change in Score
Wyoming	223	229		West Virginia	224	225	
OVERALL	224.9	226.8	+1.9*	OVERALL	219.9	224.5	+4.6*

*Significant at a $p < .05$ level

Table 3 illustrates that the states with high-stakes tests outperformed those states without high-stakes tests on the math NAEP grade 4 tests over the time period 1996-2000. However, as argued earlier, the rates by which students are excluded from the NAEP must be taken into consideration to determine whether gains and losses are clear or unclear. Using the same rules as outlined above to determine clear and unclear gains and losses, we looked at only those states for which clear gains or losses are relevant. (For 4th grade reading, Rosenshine included the following states: Arizona, Arkansas, California, Connecticut, Hawaii, Iowa, Maine, Montana, New Hampshire, Rhode Island, Utah, Washington, Wisconsin, and Wyoming. There are notable differences in the states he included and the states we included that likely came from the fact that we drew our states directly out of Table 1 of the original document.) An analysis of the data yields the following:

Table 4

Fourth grade 1996-2000 NAEP mathematics scores with states coded as clear or unclear in their gains and losses.

States without high-stakes tests:	NAEP 1996	NAEP 2000		States with high-stakes tests:	NAEP 1996	NAEP 2000	
Alaska	224	n/a		Alabama	212	218	
Arizona	218	219		California	209	214	
Arkansas	216	217		Delaware	215	n/a	
Colorado	226	n/a		Florida	216	n/a	
Connecticut	232	234		Indiana	229	234	

Georgia	216	220		Kentucky	220	221	
Hawaii	215	216		Louisiana	209	218	
Idaho	n/a	227		Maryland	220	222	
Illinois	n/a	225		Massachusetts	229	235	
Iowa	229	233		Michigan	226	231	
Kansas	n/a	232		Mississippi	208	211	
Maine	233	231		Missouri	225	229	
Minnesota	232	235		Nevada	217	220	
Montana	228	230		New Jersey	227	n/a	
Nebraska	228	226		New Mexico	214	214	
North Dakota	231	231		New York	223	227	
Oregon	224	227		North Carolina	224	232	
Rhode Island	221	225		Ohio	n/a	231	
Tennessee	219	220		Oklahoma	n/a	225	
Utah	226	227		Pennsylvania	226	n/a	
Vermont	225	232		South Carolina	213	220	
Washington	225	n/a		Texas	229	233	
Wisconsin	231	n/a	Change in Score	Virginia	222	230	Change in Score
Wyoming	223	229		West Virginia	224	225	
OVERALL	224.5	225.6	+1.1	OVERALL	210.4	215.0	+4.6*

*Significant at a p < .05 level

Compared to the reading data above, we now find the opposite when we look at the math grade 4 NAEP composite data. When states with clear effects are pulled out and analyzed, it is apparent that states with high-stakes tests are outperforming states without high-stakes tests at a statistically significant level. The scores posted by the clear states with high-stakes tests are significantly different than the scores posted by the clear states without high-stakes tests.

Again, however, what can also be drawn from Table 4 is that states with high-stakes tests are exempting more students from participating in the math grade 4 NAEP. Two times as many states with high-stakes tests exempted students and realized gains in grade 4 math achievement from 1996-2000 than did states without high-stakes tests. This, again, supports the notion that states with high-stakes tests are not all gaining in NAEP scores simply because of their high-stakes testing policies.

NAEP Math Grade 8 1996-2000

Taking Table 1 from Amrein & Berliner (2002b) and the states in which high stakes tests were implemented before 1996 and between 1996 and 2000, we re-ran our analyses using all states with high-stakes tests and, as the control group, all states without high-stakes tests for which NAEP data were available. What we found in regards to math grade 8 achievement from 1996-2000 is as follows:

Table 5

Eighth grade 1996-2000 NAEP mathematics scores (raw data)

States without high-stakes tests:	NAEP 1996	NAEP 2000	States with high-stakes tests:	NAEP 1996	NAEP 2000
Alaska	278	n/a	Alabama	256	262
Arizona	268	271	California	263	262
Arkansas	261	261	Delaware	267	n/a
Colorado	276	n/a	Florida	264	n/a
Connecticut	280	282	Indiana	275	283
Georgia	262	266	Kentucky	267	272
Hawaii	262	263	Louisiana	252	259
Idaho	n/a	278	Maryland	270	276
Illinois	n/a	277	Massachusetts	277	283
Iowa	284	n/a	Michigan	276	278
Kansas	n/a	284	Mississippi	250	254
Maine	284	284	Missouri	274	274
Minnesota	284	288	Nevada	n/a	268
Montana	283	287	New Mexico	262	260
Nebraska	283	281	New York	270	276
North Dakota	284	283	North Carolina	268	280
Oregon	277	281	Ohio	n/a	283

Rhode Island	268	273		Oklahoma	n/a	272	
Tennessee	263	263		South Carolina	260	266	
Utah	276	275		Texas	270	275	
Vermont	279	283		Virginia	270	277	
Washington	276	n/a		West Virginia	265	271	
Wisconsin	283	n/a	Change in Score				Change in Score
Wyoming	275	277					
OVERALL	275.5	276.7	+1.2*	OVERALL	266.1	271.6	+5.4*

*Significant at a p < .05 level

Table 5 illustrates the states with high-stakes tests outperformed those states without high-stakes tests on the math NAEP grade 8 1996-2000. Again, we argue that the rates by which students are excluded from the NAEP must be taken into consideration to determine whether gains and losses are clear or unclear.

Using the same rules as outlined above to determine clear and unclear gains and losses, we looked at only those states for which clear gains or losses are apparent (Note 3). An analysis of the data yields the following:

Table 6

Eighth grade 1996-2000 NAEP mathematics scores with states coded as clear or unclear in their gains and losses.

States without high-stakes tests	NAEP 1996	NAEP 2000		States with high-stakes tests	NAEP 1996	NAEP 2000	
Alaska	278	n/a		Alabama	256	262	
Arizona	268	271		California	263	262	
Arkansas	261	261		Delaware	267	n/a	
Colorado	276	n/a		Florida	264	n/a	
Connecticut	280	282		Indiana	275	283	
Georgia	262	266		Kentucky	267	272	
Hawaii	262	263		Louisiana	252	259	
Idaho	n/a	278		Maryland	270	276	

Illinois	n/a	277		Massachusetts	277	283	
Iowa	284	n/a		Michigan	276	278	
Kansas	n/a	284		Mississippi	250	254	
Maine	284	284		Missouri	274	274	
Minnesota	284	288		Nevada	n/a	268	
Montana	283	287		New Mexico	262	260	
Nebraska	283	281		New York	270	276	
North Dakota	284	283		North Carolina	268	280	
Oregon	277	281		Ohio	n/a	283	
Rhode Island	268	273		Oklahoma	n/a	272	
Tennessee	263	263		South Carolina	260	266	
Utah	276	275		Texas	270	275	
Vermont	279	283		Virginia	270	277	
Washington	276	n/a		West Virginia	265	271	
Wisconsin	283	n/a	Change in Score				Change in Score
Wyoming	275	277					
OVERALL	271.1	271.9	+0.7	OVERALL	258.8	261.8	+3.0

After the states with clear effects are pulled out and analyzed, it seems that states with high-stakes tests are outperforming states without high-stakes tests. They are not, however, outperforming states without high-stakes tests at a statistically significant level. In addition, the scores posted by the clear states with high-stakes tests are not significantly different than the scores posted by the clear states without high-stakes tests. States with high-stakes tests are not outperforming states without high-stakes tests in math grade 8 performance.

Again, what can also be drawn from Table 6 is that states with high-stakes tests are exempting more students from participating in the math grade 8 NAEP. Thirty-three percent of the states without high-stakes tests exempted more students and realized gains in math grade 8 NAEP scores. Fifty percent of the states with high-stakes tests exempted more students and realized gains in math grade 8 NAEP scores. This, again, supports our assertion that states with high-stakes tests are not gaining in NAEP scores simply because of their high-stakes testing policies.

Conclusion

In short, states with high-stakes tests seem to have outperformed states without high-stakes tests on the grade 4 math NAEP at a statistically significant level. However, gains between states with and without high stakes tests were not statistically different on the grade 4 reading or the grade 8 math NAEP. States with high-stakes tests are *not* outperforming states without high-stakes tests on both of these measures.

In addition, the rates by which personnel in states with high-stakes tests are exempting students are increasing at a faster rate than they are in states without high-stakes tests. There may be an underlying characteristic other than high-stakes tests that is causing this phenomenon, but this would take further analyses. What we do know, however, is that for the most part the gains posted by states with high-stakes tests on two of the three NAEP tests are more related to the rates by which students are exempted from the tests than they are related to high-stakes tests themselves.

We thank Professor Rosenshine for suggesting these alternative analytic techniques to us. In the end, for now, we remain unconvinced that the NAEP tests are showing much in the way of transfer effects. Given all the data we reported in our previous reports we remain unconvinced that the high-stakes tests used by states are showing systematic positive affects on audit tests used to assess transfer.

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 24, 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 24, 2003 from
http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf.

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

About the Authors

Audrey Amrein-Beardsley
Email: audrey.beardsley@cox.net

Audrey Amrein-Beardsley is a part-time Research and Evaluation Associate at a private foundation in Scottsdale, Arizona and a part-time researcher at Arizona State University. She received her PhD from Arizona State University in 2002 in Education Policy with an emphasis in Research Methodology. Her scholarly interests include the study of the intended and unintended consequences of high-stakes testing policies. To date, she has focused on the effects of high-stakes testing policies at a macro level given the frequency with which these policies are being implemented across the nation.

David C. Berliner
Regents' Professor of Education
College of Education
Arizona State University
Tempe, AZ 85287-2411

Email: berliner@asu.edu

David C. Berliner is Regents' Professor of Education at the College of Education of Arizona State University, in Tempe, AZ. He received his Ph.D. in 1968 from Stanford University in educational psychology, and has worked also at the University of Massachusetts, WestEd, and the University of Arizona. He has served as president of the American Educational Research Association (AERA), president of the Division of Educational Psychology of the American Psychological Association, and as a fellow of the Center for Advanced Study in the Behavioral Sciences and a member of the National Academy of Education. Berliner's publications include *The Manufactured Crisis*, Addison-Wesley, 1995 (with B.J. Biddle) and *The Handbook of Educational Psychology*, Macmillan, 1996 (Edited with R.C. Calfee). Special awards include the Research into Practice Award of AERA, the National Association of Secondary School Principals Distinguished Service Award, and the Medal of Honor from the University of Helsinki. His scholarly interests include research on teaching and education policy analysis.

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 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

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