1-5-2004

# Education Policy Analysis Archives 12/01

Arizona State University

University of South Florida

Education Policy Analysis Archives
Volume 12 Number 1 | January 5, 2004 | ISSN 1068-2341
http://epaa.asu.edu/epaa/v12n1/
Braun: Re-considering the Impact of High-stakes Testing

# Reconsidering the Impact of High-stakes Testing

Henry Braun
Educational Testing Service, Princeton, NJ

**Abstract**
Over the last fifteen years, many states have implemented high-stakes tests as part of an effort to strengthen accountability for schools, teachers, and students. Predictably, there has been vigorous disagreement regarding the contributions of such policies to increasing test scores and, more importantly, to improving student learning. A recent study by Amrein and Berliner (2002a) has received a great deal of media attention. Employing various databases covering the period 1990-2000, the authors conclude that there is no evidence that states that implemented high-stakes tests demonstrated improved student achievement on various external measures such as performance on the SAT®, ACT, AP®, or NAEP. In a subsequent study in which they conducted a more extensive analysis of state policies (Amrein & Berliner, 2002b), they reach a similar conclusion. However, both their methodology and their findings have been challenged by a number of authors. In this article, I undertake an extended reanalysis of one component of Amrein and Berliner (2002a). We focus on the performance of states, over the period 1992 to 2000, on the NAEP mathematics assessments for grades 4 and 8. In particular, we compare the performance of the high-stakes testing states, as designated by Amrein and Berliner, with the performance of the remaining states (conditioning, of course, on a state's participation in the relevant NAEP assessments). For each grade, when we examine the relative gains of states over the period, we find that the comparisons strongly favor the high-stakes testing states. Moreover, the results cannot be accounted for by differences between the two groups of states with respect to changes in percent of students excluded from NAEP over the same period. On the other hand, when we follow a particular cohort (grade 4, 1992 to grade 8, 1996 or grade 4, 1996 to grade 8, 2000), we find the comparisons slightly favor the low-stakes testing states, although the discrepancy can be partially accounted for by changes in the sets of states contributing to each comparison. In addition, we conduct a number of ancillary analyses to establish the robustness of our results, while acknowledging the tentative nature of any conclusions drawn from highly aggregated, observational data.

## Introduction

Since its passage in January 2002, the No Child Left Behind (NCLB) Act has already had a substantial influence on state and local education agencies as they develop accountability plans to win the approval of the U.S. Department of Education. In addition to the operational concerns of these agencies, as well as those of principals and teachers, there is considerable debate about the efficacy of externally mandated high-stakes testing in improving learning (Elmore, 2002; Lewis, 2002; Steinberg, 2003; Wolf, 2003). Indeed, most of the education and educational measurement community is doubtful that high-stakes testing will have a generally salutary effect on the quality of student learning (Linn, 2000; Mehrens, 1998), although there are contrasting views (Grissmer, Flanagan, Kawata, & Williamson, 2000). Given the level of disagreement, it is natural for both supporters and opponents to look to extant data to buttress their positions. Inasmuch as a number of states have instituted high-stakes testing policies of various kinds over the last decade or more, there is a record of results that, presumably, can yield some insights into the likely impact of such policies.

A recent, and much cited, example of this approach is the article by Amrein and Berliner (2002a). Employing some general criteria, they identify 18 states as having high-stakes testing policies and examine the achievement of their students on a number of measures, including the SAT® and the ACT, NAEP results in mathematics and reading, and Advanced Placement Program.® The rationale is that trends on the state tests cannot be relied upon as valid indicators of student learning (Linn, 2000) and, if learning is indeed taking place, then similar trends should be seen in other, related measures. Their overall conclusion was that "At the present time, there is no compelling evidence… that those policies result in transfer to the broader domains of knowledge and skill for which high-stakes test scores *must* be indicators" (p. 54).

The authors are careful to point out some of problems with each of the measures as a basis for drawing conclusions about the impact of the states' policies. NAEP results, for reasons discussed below, are perhaps the least objectionable, as well as being the most relevant to considerations related to the consequences of NCLB for elementary and middle schools. A close reading of the article, however, raises a number of methodological concerns and it is the purpose of this paper to examine those concerns through a reanalysis of the NAEP mathematics data and to explore the policy implications of the findings.

Amrein and Berliner also produced a follow-up report (Amrein & Berliner, 2002b). In that paper, they carried out a more extensive policy analysis and identified 28 states as high-stakes states. We defer discussion of this second report, as well as alternative views (e.g., Carnoy & Loeb, 2003; Raymond & Hanushek, 2003) to the Discussion section.

There is always a danger that, in carrying out these analyses, we will forget the very real limitations on the conclusions we can draw. Accordingly, we enumerate them at the outset. First, we are working with observational data so that causal inferences are not warranted. Second, these 18 states (and the other 32) have engaged in a number of education initiatives in addition to their testing policies so that ascribing differences in NAEP results (solely or principally) to the impact of high-stakes testing is problematic. A similar difficulty arises in trying to explain the results of some states in terms of (apparent) attempts to "game the system" by, for example, increasing the proportion of SD/LEP students who are excluded from participating in NAEP. (SD/LEP refers to students with disabilities and/or students with limited English proficiency—with both groups expected to perform below average.)

Perhaps the most important consideration from the policy perspective is that for all the differences among state NAEP scores and state NAEP score changes, there is much greater variability within states—and probably more to be learned from trying to understand the sources of such within state differences. For an example, see Raudenbush, Fotiu, Cheong, and Ziazi (1995). That said, the current controversy about state level results demands that we address the question in as balanced a fashion as possible.

We begin with a review the methodology of Amrein and Berliner (2002a) and then carry out a reanalysis of the cross-sectional data, followed by a reanalysis of the cohort data. Both reanalyses are repeated using the states' 25$^{th}$ percentiles (rather than the states' means) to ascertain the robustness of the results. The final section relates the findings to those in the recent literature, offers some interpretations, as well as some cautions on drawing policy conclusions from data of this type.

**Reviewing Amrein and Berliner**

Amrein and Berliner (2002a) identified 18 states as having "… the most severe consequences, that is, the highest stakes associated with their K-12 testing policies" (p. 18). All the states had regulations making high school graduation contingent on passing a high school graduation exam. They also had various combinations of other stakes relating to grade promotion contingent on examination performance, making public annual school or district report cards, as well as rewards and sanctions for schools, teachers and students (see their Table 1, p. 18). One can certainly argue with their classification that, for example, includes Kentucky and Massachusetts among the low-stakes states. In the interests of maintaining comparability, however, we have retained their classification.

The rationale for analyzing NAEP data is that NAEP is the only nationally administered achievement test—and one that students do not explicitly prepare for. Since 1990, states have had the option to participate in "State NAEP," with scores reported on the same scale as National NAEP. Consequently, states can be compared in terms of their performance on NAEP over time. If increases in state test scores are valid indicators of improved skills, then one would expect to see corresponding increases in NAEP scores.

Of course, there are some weaknesses to this approach. Student motivation to perform well on NAEP is likely not as great as it is on a high-stakes exam. (On the other hand, it is not clear why students in different states would experience differential reductions in motivation from the state test to NAEP.) States can have different policies on excluding SD/LEP students from participation in NAEP and also may differ in the extent to which the state assessment is aligned with the NAEP framework. Presumably, students in states with greater alignment might be expected to do better on NAEP than students in states with lesser alignment. On the other hand, use of SAT, ACT, or AP scores is very problematic. Aside from concerns about self-selection, it is difficult to make the case that the performance standards set by states would have had much impact on college-bound students.

With respect to the timing of policies, Amrein and Berliner (2002a) only provide the year at which the high school graduation requirement became operative in each of the 18 states. They state explicitly that "The usefulness of the NAEP analyses that follow rests on the assumption that states' other K-12 high-stakes testing policies were implemented at or around the same time as each state's high school graduation exam" (p. 36).

Their approach to the analysis of data can best be illustrated by an example. Using NAEP mathematics results for grade 4, they compute the change for the nation, and for each state, over the period 1992 to 2000. They then calculate the differential gain for each state as:

State Gain = (change for state '92 to '00) - (change for nation '92 to '00).

A positive State Gain means that over this time period the state's improvement on NAEP exceeded that of the nation. Conversely, a negative value means that the nation's improvement exceeded that of the state. It is important to recognize that in the latter case, the change for the state could be positive but just not as large as the nation's.

Using rounded values, Amrein and Berliner (2002a) find (see their Table 8) that for the eighteen high-stakes states they selected, there were 8 positive State Gains, 3 negative and 2 zeroes. There were five states where data were declared "not available." This is curious, as two states (Indiana and Minnesota) *do* have NAEP data available and we include them in our reanalysis.

Amrein and Berliner (2002a) acknowledged this result appears to support the beneficial impact of high stakes testing. However, they argue that the association between State Gain and the change in the percent of students excluded from NAEP over the same time period (r = 0.39) undercuts the interpretability of the result. When, further, they combine the analyses for 1992 to 1996 and 1996 to 2000 with the one for 1992 to 2000 (ignoring the dependency induced by the overlap in time), they reach the conclusion that "In short, when compared to the nation as a whole, high-stakes testing policies did not usually lead to improvement in the performance of students on the grade 4 NAEP math tests between 1992 and 2000" (p. 40).

In the case of grade 8 NAEP mathematics, they find (see their Table 9) that, over the period 1990–2000, five states posted gains, four losses and one remained the same. Note that eight states are missing, so these results reflect the experiences of only slightly more than half the states of interest. After aggregating results over the periods 1990 to 1992, 1992 to 1996 and 1996 to 2000 (again ignoring the overlap) and pointing to the problem of differential changes in exclusion rates, they conclude again that "there is no compelling evidence that high-stakes testing policies have improved the performance of students on the grade 8 NAEP math tests" (p. 43).

**Reanalysis**

Our approach to the question differs in a number of ways from Amrein and Berliner (2002a):

1) In addition to carrying out an analysis for the eighteen high-stakes states that were the focus of Amrein and Berliner (2002a), we carry out a parallel analysis for the other 32 states.
2) We augment our analysis by including a more comprehensive measure of states' educational reform efforts.
3) Our interpretation of the State Gain statistics is informed by consideration of the corresponding estimated standard errors. (Since the State Gain is a "difference of differences," these standard errors are not negligible, with a typical value of 2.5 points on the NAEP scale.)
4) In the analysis of the grade 8 data, we look at changes over the period 1992 to 2000, rather than 1990 to 2000. Our choice makes the analyses for grades 4 and 8 more comparable, and provides slightly more data.

The data were obtained from the National Center for Education Statistics Web site (2003). The data extracted comprise grade 4 and grade 8 NAEP mathematics results in the years 1992 and 2000 for the states and the nation (public schools only). For each jurisdiction, grade and year, we recorded the average score, the corresponding estimated standard error, and the percent of students excluded. The data are displayed in Table A1 of the appendix. We note that relevant NAEP data is available for 15 of 18 high-stakes states and 18 of 32 of the other or "low-stakes" states.

For each state and grade, we compute the State Gain and its estimated standard error. Specifically, let

$$d_4 \text{ (state)} = [\text{state('00)} - \text{state('92)}] - [\text{nat'l('00)} - \text{nat'l('92)}]$$

where the quantities on the right hand side of the equation represent the average results for grade 4. Further, for each state let

$$\text{s.e.}(d_4) = \text{(estimated) standard error of } d_4.$$

Since the four quantities contributing to $d_4$ are derived from independent samples, s.e.($d_4$) is simply the square root of the sum of the (estimated) variances of the four quantities. We also compute, for each grade and state, the changes in the percent of excluded students over the period, denoted c%ex.

Now let

$$D_4 = d_4 \,/\, \text{s.e}\,(d_4)$$

and

$$V_4 = \begin{cases} 2, \text{ if } D_4 \geq 1 \\ 1, \text{ if } 1 > D_4 \geq 0 \\ -1, \text{ if } 0 > D_4 > -1 \\ -2, \text{ if } -1 \geq D_4 \end{cases}$$

with a parallel set of definitions for $d_8$, $D_8$, and $V_8$ for the grade 8 results. Finally, we let

$$V = V_4 + V_8.$$

**Table 1**
**Basic Results for Analysis of NAEP Mathematics Scores: Grades 4 and 8 Trends for 1992 to 2000**

| State | | Policy score | $d_4$ | s.e. ($d_4$) | $\mathbf{D}_4$ | $\mathbf{V}_4$ | Changes in % excluded Gr. 4 | $d_8$ | s.e. ($d_8$) | $\mathbf{D}_8$ | $\mathbf{V}_8$ | Changes in % excluded Gr. 8 | **V** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Hi-stakes states** | **AL** | **2.20** | **1.96** | **2.45** | **0.80** | **1** | **1.27** | **2.42** | **2.75** | **0.88** | **1** | **-0.51** | **2** |
| | **GA** | **0.66** | **-3.69** | **2.05** | **-1.80** | **-2** | **1.28** | **-0.57** | **2.13** | **-0.27** | **-1** | **2.52** | **-3** |
| | **IN** | **0.90** | **5.73** | **1.95** | **2.94** | **2** | **3.46** | **5.41** | **2.24** | **2.41** | **2** | **2.72** | **4** |
| | **LA** | **-0.03** | **6.17** | **2.38** | **2.59** | **2** | **3.68** | **1.45** | **2.57** | **0.56** | **1** | **1.49** | **3** |
| | **MD** | **2.46** | **-2.66** | **2.20** | **-1.21** | **-2** | **4.88** | **3.64** | **2.30** | **1.58** | **2** | **5.88** | **0** |
| | **MN** | **-0.40** | **-0.88** | **2.04** | **-0.43** | **-1** | **2.44** | **-2.29** | **2.15** | **-1.06** | **-2** | **2.03** | **-3** |
| | **MS** | **0.55** | **1.49** | **1.97** | **0.76** | **1** | **-0.59** | **0.03** | **2.17** | **0.01** | **1** | **0.30** | **2** |
| | **NM** | **0.78** | **-7.09** | **2.42** | **-2.93** | **-2** | **4.95** | **-7.31** | **2.33** | **-3.13** | **-2** | **6.21** | **-4** |
| | **NY** | **0.09** | **0.46** | **2.21** | **0.21** | **1** | **6.24** | **2.29** | **3.21** | **0.71** | **1** | **4.64** | **2** |
| | **NC** | **1.60** | **11.92** | **1.94** | **6.14** | **2** | **9.48** | **14.18** | **2.07** | **6.86** | **2** | **10.59** | **4** |
| | **OH** | **1.15** | **4.21** | **2.17** | **1.94** | **2** | **3.91** | **7.00** | **2.47** | **2.83** | **2** | **2.61** | **4** |
| | **SC** | **0.90** | **0.27** | **2.16** | **0.12** | **1** | **2.65** | **-1.96** | **2.12** | **-0.93** | **-1** | **0.94** | **0** |
| | **TN** | **0.32** | **1.23** | **2.37** | **0.52** | **1** | **-0.10** | **-2.94** | **2.55** | **-1.15** | **-2** | **-0.33** | **-1** |
| | **TX** | **-0.66** | **7.09** | **2.12** | **3.34** | **2** | **7.86** | **2.71** | **2.34** | **1.16** | **2** | **2.99** | **4** |
| | **VA** | **0.55** | **1.98** | **2.21** | **0.89** | **1** | **5.55** | **1.27** | **2.29** | **0.55** | **1** | **4.69** | **2** |
| Lo-stakes states | AZ | -0.40 | -4.14 | 2.17 | -1.91 | -2 | 6.92 | -2.20 | 2.36 | -0.93 | -1 | 3.37 | -3 |
| | AR | -0.27 | -0.80 | 1.91 | -0.42 | -1 | 1.15 | -2.49 | 2.21 | -1.13 | -2 | 1.94 | -3 |
| | CA | 0.09 | -2.49 | 2.72 | -0.91 | -1 | -3.24 | -6.27 | 2.92 | -2.14 | -2 | 0.42 | -3 |
| | CT | 1.29 | -0.21 | 2.05 | -0.10 | -1 | 3.37 | 0.62 | 2.19 | 0.28 | 1 | 3.60 | 0 |
| | HI | 0.32 | -5.86 | 2.14 | -2.73 | -2 | 4.42 | -2.18 | 2.04 | -1.07 | -2 | 2.42 | -4 |
| | ID | -0.27 | -2.32 | 1.99 | -1.17 | -2 | 2.46 | -4.72 | 1.97 | -2.39 | -2 | 1.59 | -4 |
| | KY | 1.97 | -1.71 | 1.99 | -0.86 | -1 | 4.95 | 1.77 | 2.20 | 0.81 | 1 | 4.91 | 0 |
| | ME | 1.29 | -8.73 | 1.85 | -4.72 | -2 | 4.46 | -2.54 | 2.00 | -1.27 | -2 | 4.13 | -4 |
| | MA | 0.32 | 0.71 | 2.05 | 0.34 | 1 | 3.45 | 2.79 | 2.07 | 1.35 | 2 | 3.99 | 3 |
| | MI | 0.43 | 3.35 | 2.56 | 1.31 | 2 | 3.11 | 3.55 | 2.47 | 1.44 | 2 | 0.48 | 4 |
| | MO | 1.02 | -1.32 | 2.10 | -0.63 | -1 | 5.34 | -5.10 | 2.28 | -2.24 | -2 | 4.11 | -3 |
| | NE | -1.61 | -7.04 | 2.46 | -2.87 | -2 | 3.45 | -4.58 | 2.02 | -2.26 | -2 | -0.57 | -4 |
| | ND | -0.03 | -5.42 | 1.71 | -3.18 | -2 | 3.98 | -7.69 | 2.02 | -3.81 | -2 | 1.40 | -4 |
| | OK | 0.43 | -2.93 | 2.03 | -1.45 | -2 | 3.18 | -4.03 | 2.27 | -1.78 | -2 | 2.31 | -4 |
| | RI | 0.09 | 1.53 | 2.32 | 0.66 | 1 | 5.89 | -0.03 | 1.84 | -0.01 | -1 | 6.67 | 0 |
| | UT | 1.15 | -4.40 | 2.00 | -2.20 | -2 | 2.63 | -6.45 | 1.87 | -3.45 | -2 | 1.45 | -4 |
| | WV | 0.90 | 1.92 | 2.03 | 0.95 | 1 | 5.65 | 4.15 | 1.91 | 2.17 | 2 | 5.29 | 3 |
| | WY | -0.95 | -3.78 | 2.03 | -1.86 | -2 | 2.55 | -5.94 | 1.93 | -3.07 | -2 | 0.01 | -4 |

Table 1 displays the relevant quantities. (Note: The policy score will be defined presently.) We observe that for high-stakes states, $d_4$ ranges from –7.09 to 11.92, with a median of 1.49 and a mean of 1.88; $d_8$ ranges from –7.31 to 14.18, with a median of 1.45 and a mean of 1.69. For low-stakes states, $d_4$ ranges from –8.73 to 3.35, with a median of –2.41 and a mean of –2.42; $d_8$ ranges from –7.69 to 4.15, with a median of –2.52 and a mean of –2.30. Thus, we see that the typical State Gain for high-stakes states is substantially larger than the typical State Gain for low-stakes states in both grades 4 and 8.

At grade 4, the difference in means between the high-stakes and low-stakes states is 4.3 score points and at grade 8 it is 3.99 score points. Note that in computing the difference in means, the gain of the nation over the period 1992 to 2000 is eliminated. Consequently, such differences provide a direct comparison between the typical gains for high-stakes states and low-stakes states. Some might prefer such comparisons because the results for the nation are influenced by all the states we are considering, as well as the states that did not participate in both NAEP administrations. However, we have chosen to follow the approach of Amrein and Berliner (2002a) in order to facilitate comparisons between our results and theirs.

While there certainly is interest in the State Gains (*d*s) themselves, we believe there is also value in comparing states in terms of the *V*s, which are essentially discretized effect sizes. Specifically, $V_k$ ($k$ = 4 or 8) gives a state 2 “credits” if $d_k$ exceeds one standard error (in one direction or the other). While the usual criterion for statistical significance (which is not particularly appropriate in this setting) would require exceeding two standard errors, there is practical interest in identifying states whose relative gain is at least greater than one standard error—given the magnitude of the standard errors, the level of dispersion in the $d_k$s among the states, and the fact that the national gain (although statistically independent of the state gains) is influenced by the educational policies of the various states.

A state that presents what one might term a strongly consistent picture of relative improvement over the nation (i.e., $D_4 > 1$ and $D_8 > 1$) is awarded 4 credits. One that presents a moderately consistent picture (e.g., $D_4 > 1$ and $1 > D_8 > 0$) is awarded 3 credits and one that presents a mildly consistent picture (i.e., $1 > D_4 > 0$ and $1 > D_8 > 0$) is awarded 2 credits. Note that this coding scheme limits the influence of outliers and allows us to distinguish most configurations of $D_4$ and $D_8$.

The distributions of *V* for the high-stakes testing states and the remaining states are presented in Table 2. For the first group, we have values of *V* for 15 out of the 18 states, while for the second group we have values of *V* for 18 out of 32 states. There is a striking difference between the two groups of states: High-stakes states are more likely to show strongly consistent improvement relative to the nation (*V* = 4) than low-stakes states (4/15 vs. 1/18) and much less likely to show strongly consistent lack of improvement (*V* = –4) relative to the nation (1/15 vs. 8/18). The story remains qualitatively the same if we compare the groups with less stringent cut-offs.

**Table 2**
***Distribution of V for Hi-stakes and Lo-stakes States***

| # Lo-stakes states Total = 18 | V | # Hi-stakes states Total = 15 |
|---|---|---|
| 1 | **4** | 4 |
| 2 | **3** | 1 |
| 0 | **2** | 4 |
| 0 | **1** | 0 |
| 3 | **0** | 2 |
| 0 | **-1** | 1 |
| 0 | **-2** | 0 |

| | | |
|---|---|---|
| 4 | **-3** | 2 |
| 8 | **-4** | 1 |

In summary, high-stakes testing states that participated in the NAEP mathematics assessment in both 1992 and 2000 typically showed improvement relative to the nation while low-stakes testing states that participated in the NAEP mathematics assessment in both 1992 and 2000 typically showed lack of improvement relative to the nation. (We must be careful to condition on participation in NAEP since a large number of low-stakes states did not participate in NAEP in one or both of the years under study.) The question is how to interpret the comparison.

With respect to the results for the high-stakes states, Amrein and Berliner (2002a) discount the finding, in part, because of the empirical association between State Gain and the change in percent excluded. This is a reasonable argument but one that deserves further scrutiny for at least two reasons. First, the observed correlation may be unduly influenced by an outlying observation and, second, there are other, observable and unobservable characteristics of states that may also account for some of the differences among states. (It also should be noted that the 1992 exclusion rates are not strictly comparable to those in 1996 and 2000. The former were calculated as an average over mathematics and reading, while the latter two are reported for mathematics only.)

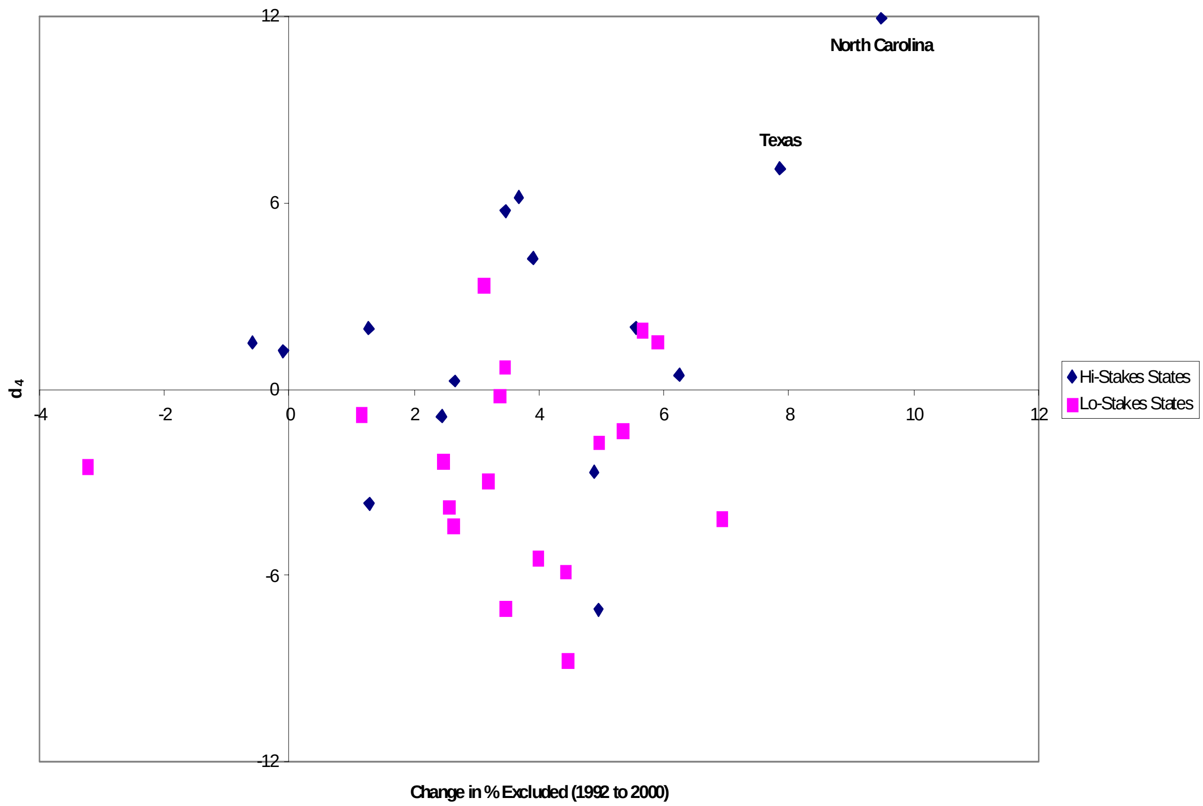


***Figure 1a.*** **Grade 4: $d_4$ vs. Change in % Excluded (1992 to 2000).**

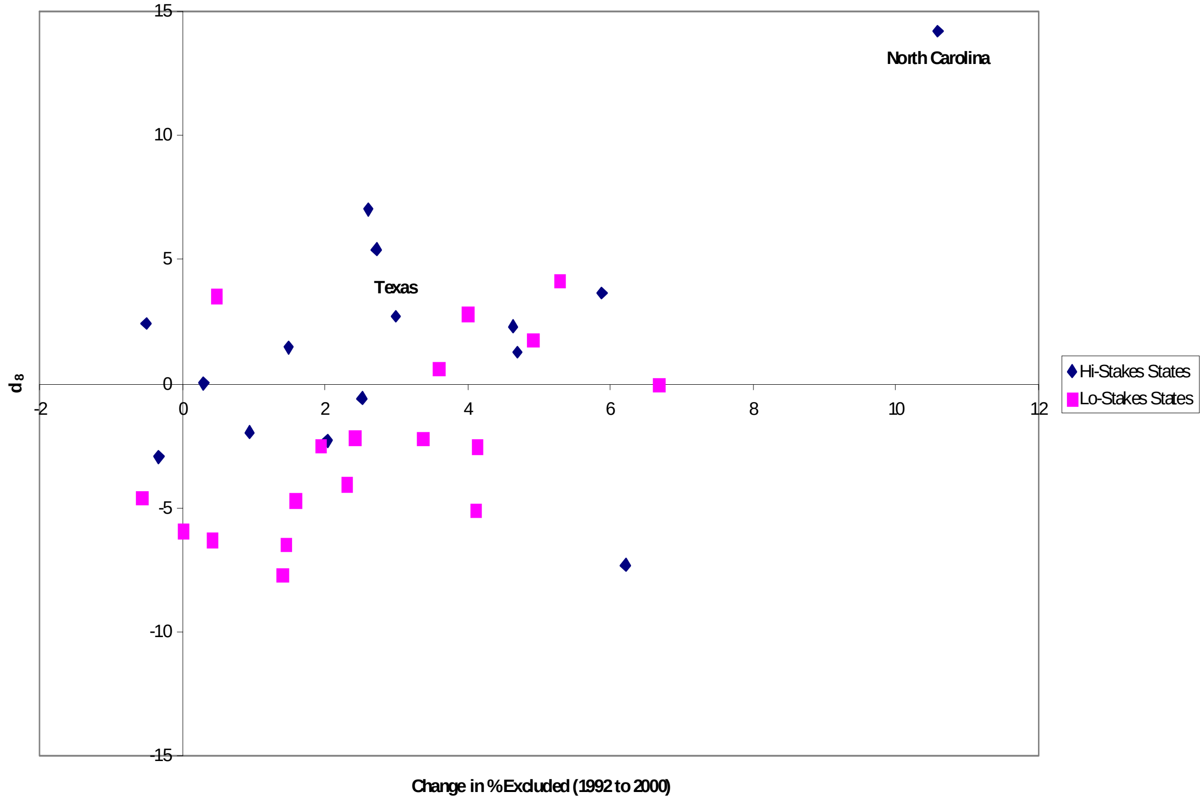


***Figure 1b.*** **Grade 8: $d_8$ vs. Change in % Excluded (1992 to 2000).**

Figure 1a displays a plot of $d_4$ against c%ex and Figure 1b displays a plot of $d_8$ against c%ex. North Carolina is a clear outlier on both plots, while Texas is an outlier in Figure 1a. Table 3 presents the correlations for the two groups of states, including the case for the high-stakes states with North Carolina removed. In the fourth grade, we see that the correlation for the high-stakes states is indeed substantial, but markedly reduced when North Carolina is deleted. For the eighth grade, the reduction is even more dramatic. On the other hand, for low stakes states in the eighth grade, the correlation is quite high. One might, therefore, plausibly argue that the results for the low-stakes states would be further depressed (relative to those for the high-stakes states) if their apparent relationship to c%ex were somehow taken into account.

**Table 3**
***Correlations Between State Gains and Change in % Excluded for Years 1992 to 2000***

| | State gains | |
|---|---|---|
| | Grade 4 ($d_4$) | Grade 8 ($d_8$) |
| Hi-stakes (# = 15) | 0.44 | 0.49 |
| Hi-stakes w/o NC (# = 14) | 0.17 | -0.01 |
| Lo-stakes (# = 18) | 0.02 | 0.55 |

It is often the case that gain scores are negatively correlated with the base year score. Accordingly, in Figures 2a and 2b we plot $d_4$ against the grade 4 state score ('92) and $d_8$ against the grade 8 state score ('92). In both cases, we observe the expected negative correlations. Plots of $d_4$ and $d_8$ against their standard errors were not informative and are not presented.

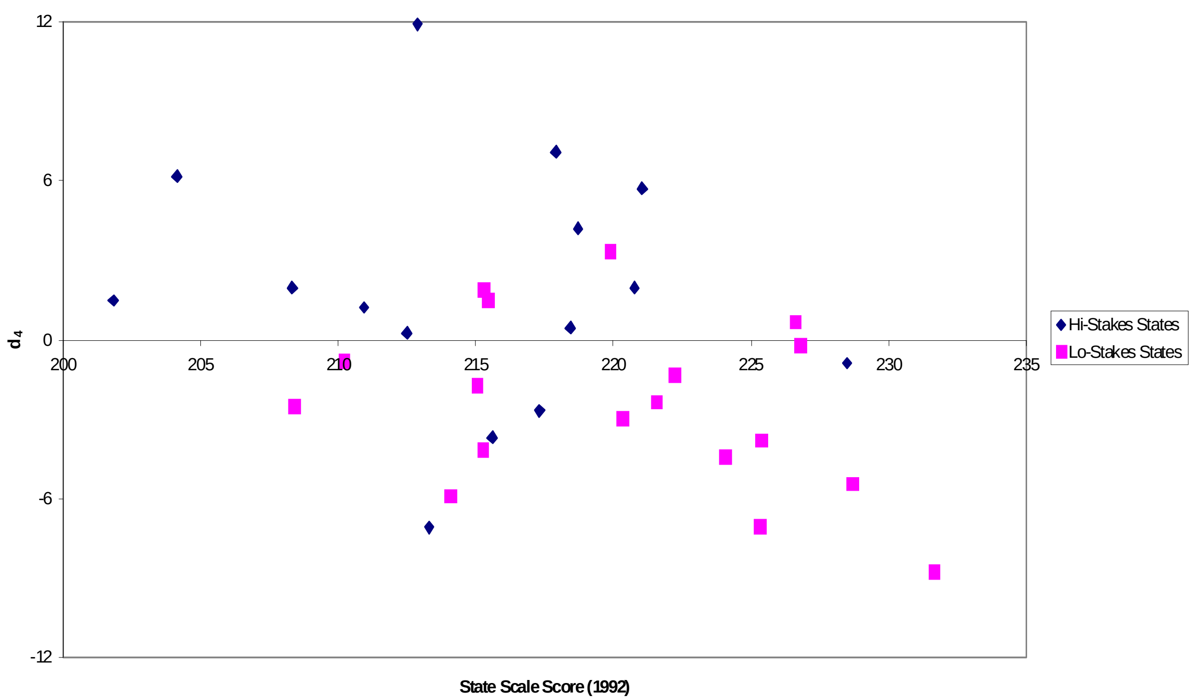


***Figure 2a.*** **Grade 4: $d_4$ vs. State Scale Score (1992).**

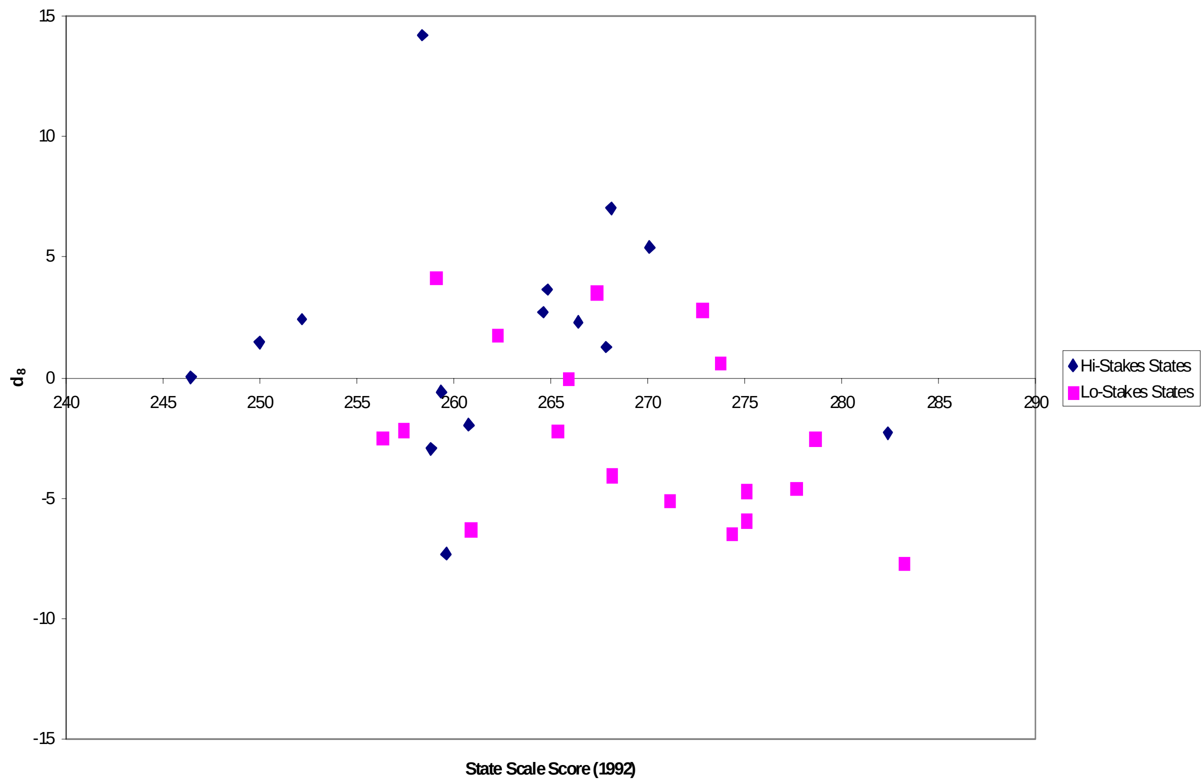


***Figure 2b.*** **Grade 8: $d_8$ vs. State Scale Score (1992).**

As was noted in the Introduction, changes in state NAEP scores over time can be the result of many factors in addition to percent of students excluded and testing policies. In particular, other educational interventions that have been adopted by the state with the intention of raising the academic achievement of its students may well have their intended effect, at least to some degree. It would be helpful, therefore, to have a broader measure of each state's educational policy efforts to incorporate into an explanatory framework. Fortunately, one such a measure has been formulated and quantified as part of a study of the influence of standards-based reform on changes in classroom practice (Swanson & Stevenson, 2002).

Drawing on studies conducted by the Council of Chief State School Officers, Swanson and Stevenson graded each state on each of 22 policy activities organized into four categories: content standards, performance standards, aligned assessments and professional standards. Grades were assigned on a three point scale: does not have such a policy (0), is developing one (1), or has enacted such a policy as of 1996 (2). They then carried out a Rasch analysis using this 50 x 22 data array, yielding a "state (policy) activism score" for each state. They report a low level of item misfit. For more details, consult their article.

In view of the comprehensiveness of the policy information employed and that 1996 falls in the middle of the period of interest, we propose to use the policy activism scale as another possible explanatory variable in our effort to account for differences among states in State Gains.

The policy activism scores are located in the second column of Table 1. Figures 3a and 3b display plots of $d_4$ and $d_8$ against activism scores.

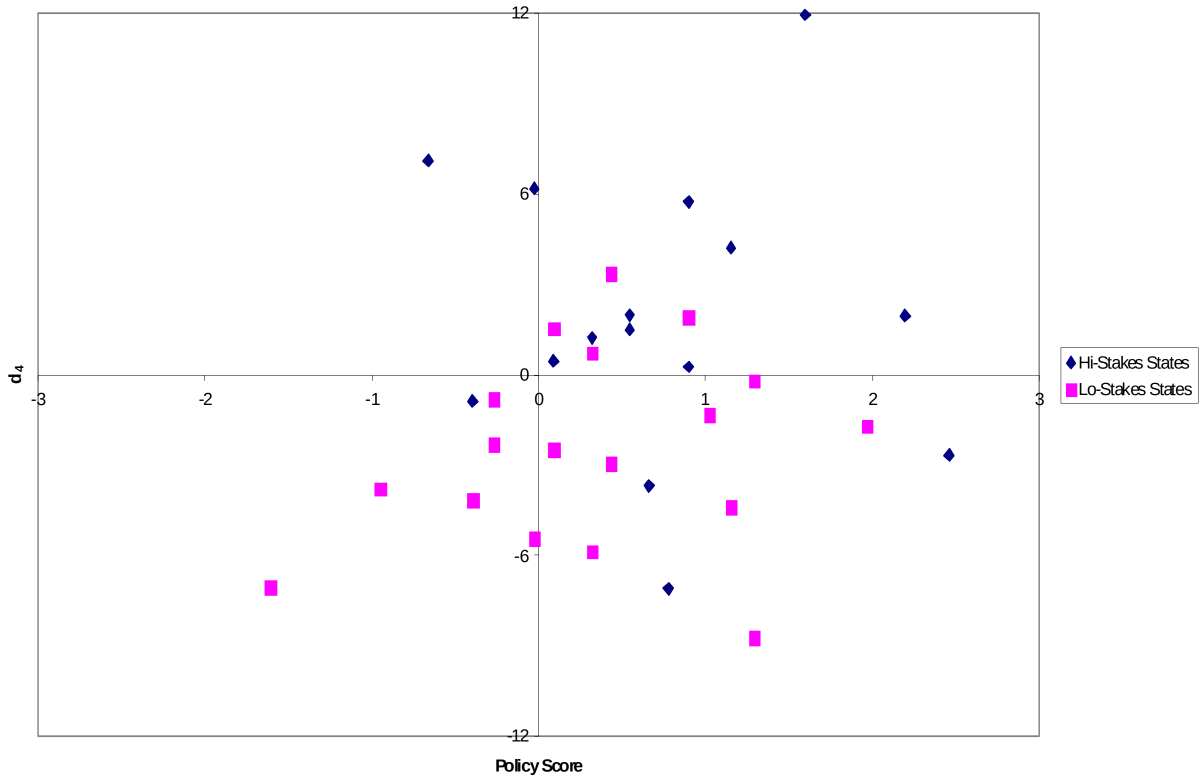


***Figure 3a.*** **Grade 4:** $d_4$ **vs. Policy Score.**

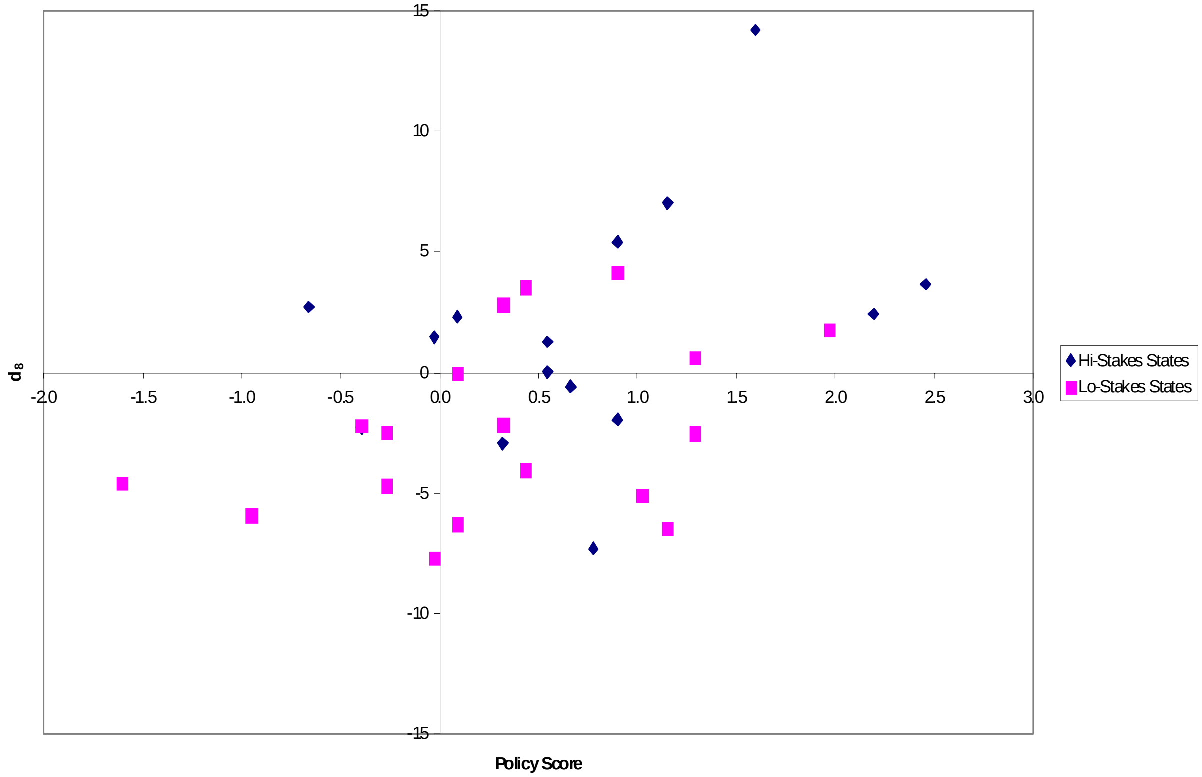


***Figure 3b.*** **Grade 8: $d_8$ vs. Policy Score.**

Since the mean policy score over the 50 states is 0.28, we note that the 33 states that we are examining tend to have scores above the mean. The median for the high-stakes states is 0.66 and the median for the low-stakes states is 0.32. Correlations between State Gain and policy scores for the two groups of states are presented in Table 4. We observe the relationship is moderately strong and positive in grade 8, but rather mixed in grade 4. Again, North Carolina exerts considerable leverage on the results for the high-stakes states.

**Table 4**
***Correlations Between State Gains for Years 1992 to 2000 and Policy Score***

| | State gains | |
|---|---|---|
| | Grade 4 ($d_4$) | Grade 8 ($d_8$) |
| Hi-stakes (# = 15) | -0.07 | 0.37 |
| Hi-stakes w/o NC (# = 14) | -0.29 | 0.26 |
| Lo-stakes (# = 18) | 0.22 | 0.38 |

Before proceeding to the next stage of the comparison between the high-stakes and low-stakes states, it might be of interest to compare the V distributions of high activism and low activism states, defined by whether they are above or below the mean policy score of 0.28, respectively. The results are presented in Table 5, which is analogous to Table 2. This comparison involves 21 out 27 high activism states and 12 out of 23 low activism states. While the comparison favors the high activism states, it is less clear-cut than the one in Table 2. Note that the V values of the high activism states fall about equally above and below zero. On the other hand, the V values of the low activism states are more likely to be negative. Thus, somewhat surprisingly, the categorization employed in Amrein and Berliner (2002a) seems to provide a sharper contrast than the categorization based on the broader policy analysis employed by Swanson and Stevenson (2002).

**Table 5**
***Distribution of V for Hi-policy Score and Lo-policy Score States***

| # Lo-policy score states Total = 12 | V | # Hi-policy score states Total = 21 |
|---|---|---|
| 1 | **4** | 4 |
| 1 | **3** | 2 |
| 1 | **2** | 3 |
| 0 | **1** | 0 |
| 1 | **0** | 4 |
| 0 | **-1** | 1 |
| 0 | **-2** | 0 |
| 4 | **-3** | 2 |
| 4 | **-4** | 5 |

Returning to the main thread of our reanalysis, we carry out a multiple regression of $d_4$ on three explanatory variables: state score ('92), c%ex and activism score, and an analogous regression for $d_8$. In both regressions, we leave out North Carolina and Texas because they are outliers in one or both panels of Figure 1. The essential elements of the regression output are presented in Tables 6a and 6b.

**Table 6a**

***Grade 4: Regression of $d_4$ on Policy Score, 1992 State Scale Score, Change in % Excluded for Years 1992 to 2000***

| ANOVA | | | | | |
|---|---|---|---|---|---|
| | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 3 | 67.1 | 22.4 | 1.7 | 0.2 |
| Residual | 27 | 350.3 | 13.0 | | |
| Total | 30 | 417.4 | | | |

| | *Coefficients* | *Standard error* | *t stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 40.6 | 20.9 | 1.9 | 0.1 |
| Policy score | 0.5 | 0.8 | 0.6 | 0.5 |
| Score | -0.2 | 0.1 | -2.0 | 0.1 |
| Change in % excluded | 0.1 | 0.3 | 0.3 | 0.8 |

| SUMMARY OUTPUT | |
|---|---|
| *Regression statistics* | |
| Multiple R | 0.40 |
| R square | 0.16 |
| Adjusted R square | 0.07 |
| Standard error | 3.60 |
| Observations | 31.00 |

**Table 6b**

***Grade 8: Regression of*** $d_8$ ***on Policy Score, 1992 State Scale Score, Change in % Excluded for Years 1992 to 2000***

| ANOVA | | | | | |
|---|---|---|---|---|---|
| | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 3 | 113.3 | 37.8 | 2.9 | 0.1 |
| Residual | 27 | 353.8 | 13.1 | | |
| Total | 30 | 467.2 | | | |

| | *Coefficients* | *Standard error* | *t stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 21.3 | 20.2 | 1.1 | 0.3 |
| Policy score | 1.5 | 0.9 | 1.7 | 0.1 |
| Score | -0.1 | 0.1 | -1.2 | 0.2 |
| Change in % excluded | 0.3 | 0.3 | 0.9 | 0.4 |

| SUMMARY OUTPUT | |
|---|---|
| *Regression statistics* | |
| Multiple R | 0.49 |
| R square | 0.24 |
| Adjusted R square | 0.16 |
| Standard error | 3.62 |
| Observations | 31.00 |

For Grade 4, the $R^2 = 0.16$ (adjusted $R^2 = 0.07$) so clearly the three explanatory variables do not account for very much of between-state variation; only state score ('92) is marginally significant. Overall, residual plots against each of the explanatory variables do not reveal any patterns. However, the residuals for the 13 high-stakes states (i.e. not including Texas and North Carolina) tend be more positive than the residuals for the 18 low-stakes states. This is to be expected given the results in Tables 1, 2 and 6.

Figure 4a presents the residual plot against c%ex. The residuals for Texas and North Carolina were obtained by substituting their values for the three explanatory variables into the regression equation presented in Table 6a (which was estimated using the other 31 states). We note that Texas and North Carolina are outliers in the sense that they have both the largest values on c%ex and the largest positive residuals. On the other hand, for the other states there appears to be no association (linear or otherwise) between c%ex and state gain.

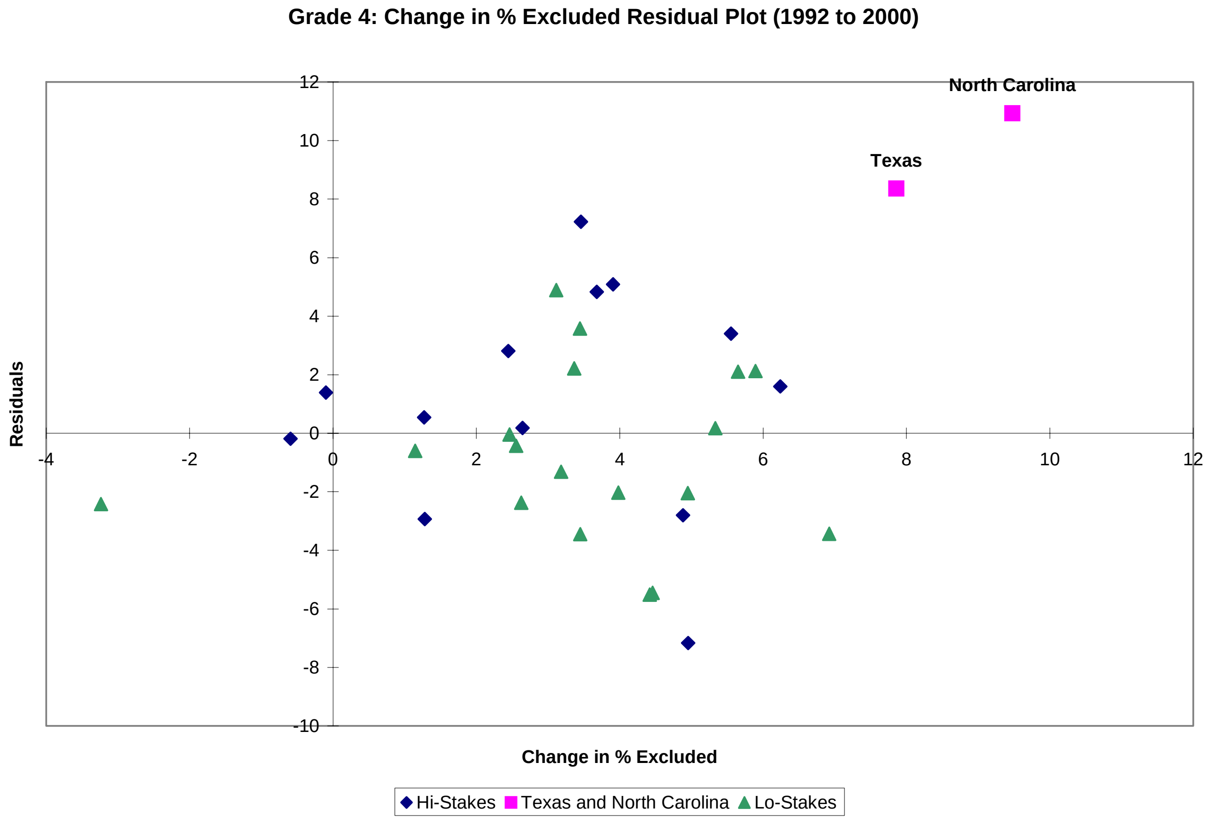


***Figure 4a.*** **Grade 4: Plot of residuals vs. Change in % Excluded (1992 to 2000). Residuals obtained from a regression of $d_4$ on state score ('92), c%ex and policy score.**

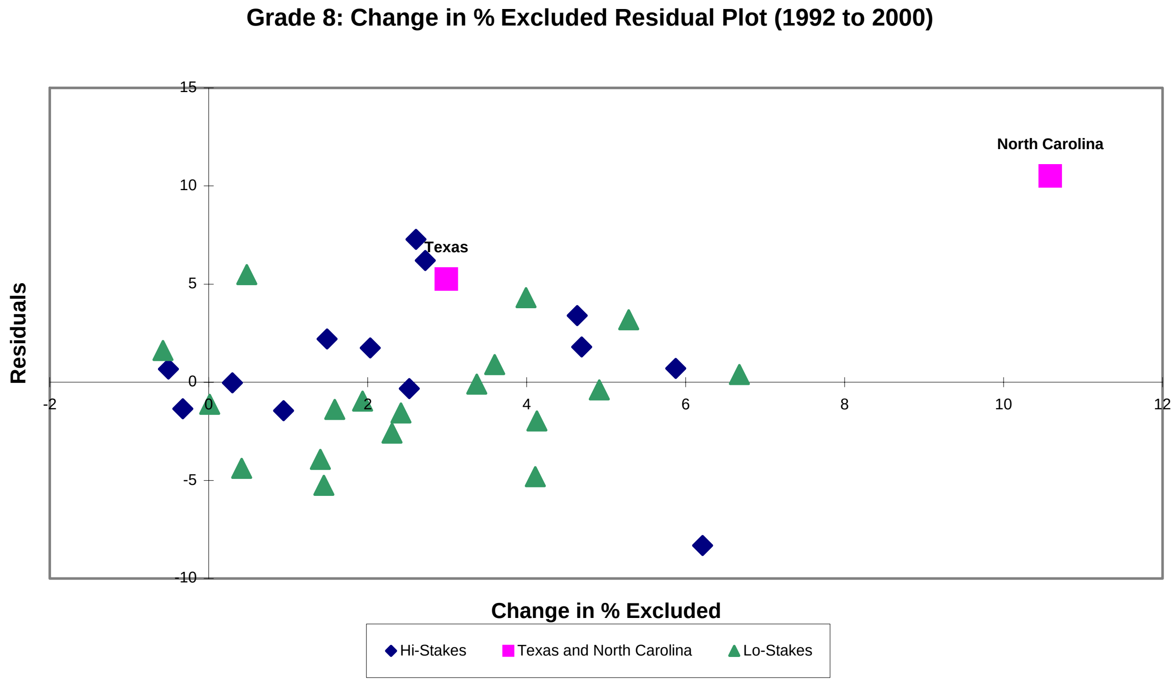


***Figure 4b.*** **Grade 8: Plot of residuals vs. Change in % Excluded (1992 to 2000). Residuals obtained from a regression of $d_8$ on state score ('92), c%ex and policy score.**

Turning to Grade 8 (Table 6b), we note that the $R^2 = 0.24$ (adjusted $R^2 = 0.16$) and that the only explanatory variable that approaches significance is policy score. Overall, the residual plots again reveal no interesting patterns, except that high-stakes states tend to have more positive residuals than do low-stakes states. Figure 4b presents the residual plot against c%ex, with the residuals for Texas and North Carolina added. North Carolina remains an outlier, but not Texas. For the other states, there does not appear to be an association between c%ex and state gain.

In view of the above analysis, it is not appropriate to discount the differences in results between the high-stakes and low-stakes states (e.g. Table 2) by arguing they are strongly influenced by differences in changes in percent of students excluded over the period 1992 to 2000. That argument is simply not supported by the data.

One might want to distinguish the results for North Carolina from those of the other states, arguing that the unusually large value of c%ex "explains" the unusually large value of State Gain. If that were the case, then school officials in North Carolina would have been much more adept than officials in other states in excluding SD/LEP students who would have done poorly on NAEP. In particular, school officials in New Mexico, which also experienced a large increase in percent of students excluded (particularly in Grade 8) but large negative State Gains, would have much to learn from their counterparts in North Carolina! A more circumspect statement about North Carolina is that its State Gain may well be a consequence of both its reform policies and the increase in excluded students—but that with the data available we are neither able to determine the relative contributions of these two factors nor those of other factors.

**Cohort Analyses**

Amrein and Berliner (2002a) correctly point out that a weakness of the repeated cross-sectional studies described above is that real changes over time in student test performance are confounded with changes in the characteristics of successive cohorts that are unrelated to school effects but associated with performance. For example, in a particular state, grade 4 students in 2000 might be more disadvantaged than were grade 4 students in 1992 and, therefore, perform more poorly on NAEP even if the productivity of the state's schools remained unchanged.

The structure of the NAEP system makes possible another way of looking at a state's performance. Since NAEP tested students in mathematics in both grades 4 and 8 every four years, we can determine the gains of the cohort tested in grade 4 in 1992 and again in grade 8 in 1996, as well as the gains of the cohort tested in grade 4 in 1996 and again in grade 8 in 2000. Although the actual students tested four years apart are not the same students (i.e., this is not a true longitudinal study like High School and Beyond), each group is a probability sample of their respective cohorts. Thus, the observed gain is an approximately unbiased estimate of the population gain over the period in question. The word "approximately" is appropriate since there are inflows and outflows over the four years, as well as differential rates of exclusions and non-response at school and student levels. Nonetheless, the results should be sufficiently accurate for our purposes.

Others have also studied cohort gains and obtained results that cast a different light on between state comparisons. Examining data for 1992 and 1996, Barton and Coley (1998) concluded that "Most of the states are not significantly different from each other in terms of cohort growth from the fourth to the eighth grade." They point out, for example, Maine ranks near the top for grade 4 in 1992 and for grade 8 in 1996, while Arkansas ranks near the bottom in both years. Nevertheless, both cohorts gained 52 points over the four-year period.

We now carry out an analysis that parallels the one described in the previous section. The data extracted from the NCES Web site comprise grade 4 NAEP mathematics results for 1992 and 1996 and grade 8 NAEP mathematics results for 1996 and 2000, for the states and the nation (public schools only). For each jurisdiction, for the indicated grade and year, we recorded the average score, the corresponding estimated standard error, and the percent of students excluded. The data are displayed in Table A2 of the appendix.

For each state and grade, we compute the State Cohort Gain (1992 to 1996) as

$$g_1 = [\text{state(grade 8, 1996)} - \text{state(grade 4, 1992)}] - [\text{national (grade 8, 1996)} - \text{national (grade 4, 1992)}]$$

where the quantities on the right hand side of the equation represent the average results for the indicated grade and year. Further, for each state let

$$\text{s.e.}\,(g_1) = \text{(estimated) standard error of } g_1.$$

As before, since the four quantities contributing to $g_1$ are derived from independent samples, $\text{s.e.}(g_1)$ is the square root of the sum of their (estimated) variances. We also computed the changes from 1992 to 1996 in the percent of excluded students in the cohort. Now let

$$G_1 = g_1 / \text{s.e.}\,(g_1)$$

and

$$W_1 = \begin{cases} 2, \text{ if } G_1 \geq 1 \\ 1, \text{ if } 1 > G_1 \geq 0 \\ -1, \text{ if } 0 > G_1 \geq -1 \\ -2, \text{ if } -1 > G_1 \end{cases}$$

There is a set of analogous definitions for $g_2$, $G_2$, and $W_2$ based on the cohort gains from grade 4 in 1996 to grade 8 in 2000. Finally, we let

$$W = W_1 + W_2.$$

Table 7 displays the relevant quantities. For high-stakes states, $g_1$ ranges from –5.06 to 3.63, with a median of –1.98 and a mean of -1.18. For low-stakes states, $g_1$ ranges from –3.86 to 5.51, with a median of 0.75 and a mean of 0.68. Turning to the second cohort, for high-stakes states, $g_2$ ranges from –7.81 to 3.73, with a median of –1.20 and a mean of –1.08. For low-stakes states, $g_2$ ranges from –6.56 to 7.00, with a median of 0.12 and a mean of 0.06.

Thus, the difference in means for the earlier cohort between high-stakes and low-stakes states is $-1.86$ and for the later cohort the difference is –1.14. As before, the growth of the nation over the relevant four-year period is eliminated when we consider these differences in means. Interestingly, the results for low-stakes states are now somewhat better than those for high-stakes states—a reversal of what we found when we looked at change over time in a particular grade.

Note also that $W_1$ and $W_2$ are based on independent samples, so that W (when it is defined) is a reasonable choice as a summary measure of the state's relative performance over the period 1992 to 2000. On the other hand, there is value in studying $W_1$ and $W_2$ separately, to see if there are any trends over time and to examine patterns of association with c%ex and policy score.

**Table 7**
***Basic Results for Cohort Analysis of NAEP Mathematics Scores***

| | | | Chohort 1992 to 1996 | | | | | Cohort 1996 to 2000 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| State | | Policy score | $g_1$ | s.e. ( $g_1$ ) | $G_1$ | $W_1$ | Changes in % excluded | $g_2$ | s.e. ( $g_2$ ) | $G_2$ | $W_2$ | Changes in % excluded | W |
| Hi-stakes states | AL | 2.20 | -3.66 | 3.03 | -1.21 | -2 | 2.64 | -1.56 | 2.54 | -0.62 | -1 | -1.34 | -3 |
| | FL | -0.27 | -1.98 | 2.78 | -0.71 | -1 | 1.56 | *** | *** | *** | *** | *** | *** |
| | GA | 0.66 | -5.06 | 2.52 | -2.01 | -2 | 1.76 | -1.20 | 2.36 | -0.51 | -1 | -0.03 | -3 |
| | IN | 0.90 | 2.56 | 2.29 | 1.12 | 2 | 2.29 | 1.58 | 2.23 | 0.71 | 1 | 2.02 | 3 |
| | LA | -0.03 | -3.69 | 2.59 | -1.42 | -2 | 1.96 | -2.11 | 2.29 | -0.92 | -1 | -1.90 | -3 |
| | MD | 2.46 | 0.43 | 2.88 | 0.15 | 1 | 2.62 | 3.25 | 2.50 | 1.30 | 2 | 2.88 | 3 |
| | MN | -0.40 | 3.63 | 2.17 | 1.67 | 2 | -0.40 | 3.39 | 2.24 | 1.51 | 2 | -0.63 | 4 |
| | MS | 0.55 | -3.54 | 2.17 | -1.64 | -2 | 1.87 | -6.47 | 2.22 | -2.91 | -2 | 1.50 | -4 |
| | NV | 0.32 | *** | *** | *** | *** | *** | -1.51 | 2.08 | -0.73 | -1 | 1.27 | *** |
| | NM | 0.78 | -3.26 | 2.38 | -1.37 | -2 | 0.45 | -6.07 | 2.80 | -2.17 | -2 | -0.39 | -4 |
| | NY | 0.09 | -0.14 | 2.54 | -0.06 | -1 | 2.34 | 1.56 | 2.77 | 0.56 | 1 | 5.33 | 0 |
| | NC | 1.60 | 3.02 | 2.31 | 1.31 | 2 | 0.59 | 3.73 | 2.11 | 1.77 | 2 | 6.96 | 4 |
| | SC | 0.90 | -3.65 | 2.38 | -1.54 | -2 | 0.98 | 1.09 | 2.32 | 0.47 | 1 | 1.20 | -1 |
| | TN | 0.32 | 0.24 | 2.43 | 0.10 | 1 | 0.46 | -7.81 | 2.58 | -3.02 | -2 | -1.84 | -1 |
| | TX | -0.66 | 0.35 | 2.37 | 0.15 | 1 | 1.06 | -5.94 | 2.41 | -2.47 | -2 | -0.75 | -1 |
| | VA | 0.55 | -2.94 | 2.50 | -1.18 | -2 | 2.02 | 1.96 | 2.42 | 0.81 | 1 | 3.29 | -1 |
| Lo-stakes states | AZ | -0.40 | 0.69 | 2.38 | 0.29 | 1 | 3.59 | 1.07 | 2.66 | 0.40 | 1 | -3.34 | 2 |
| | AR | -0.27 | -0.48 | 2.28 | -0.21 | -1 | 1.58 | -6.56 | 2.40 | -2.73 | -2 | 1.49 | -3 |
| | CA | 0.09 | 2.44 | 2.82 | 0.87 | 1 | -2.21 | 0.97 | 3.05 | 0.32 | 1 | -7.07 | 2 |
| | CO | 0.66 | 2.66 | 2.06 | 1.29 | 2 | -0.83 | *** | *** | *** | *** | *** | *** |
| | CT | 1.29 | 0.86 | 2.15 | 0.40 | 1 | 1.70 | -2.20 | 2.20 | -1.00 | -1 | 2.10 | 0 |
| | DE | 0.21 | -3.10 | 1.90 | -1.63 | -2 | 3.34 | *** | *** | *** | *** | *** | *** |
| | HI | 0.32 | -3.86 | 2.18 | -1.77 | -2 | -0.54 | -4.27 | 2.38 | -1.79 | -2 | 1.50 | -4 |
| | IA | -1.61 | 2.17 | 2.20 | 0.99 | 1 | 1.97 | *** | *** | *** | *** | *** | *** |
| | KY | 1.97 | -0.39 | 2.06 | -0.19 | -1 | 1.40 | -0.50 | 2.21 | -0.23 | -1 | 3.75 | -2 |
| | ME | 1.29 | 0.49 | 2.18 | 0.23 | 1 | -0.95 | -0.64 | 2.06 | -0.31 | -1 | 1.04 | 0 |
| | MA | 0.32 | -0.96 | 2.55 | -0.38 | -1 | 1.03 | 2.08 | 2.27 | 0.92 | 1 | 3.11 | 0 |
| | MI | 0.43 | 5.06 | 2.87 | 1.76 | 2 | -0.14 | 0.12 | 2.44 | 0.05 | 1 | 0.37 | 3 |
| | MO | 1.02 | -0.86 | 2.33 | -0.37 | -1 | 2.73 | -3.23 | 2.24 | -1.44 | -2 | 3.64 | -3 |
| | MT | -1.26 | *** | *** | *** | *** | *** | 7.00 | 2.18 | 3.20 | 2 | 0.67 | *** |
| | NE | -1.61 | 5.51 | 2.16 | 2.55 | 2 | 0.20 | 1.00 | 2.10 | 0.48 | 1 | -1.52 | 3 |
| | ND | -0.03 | 3.63 | 1.88 | 1.93 | 2 | 1.58 | 0.10 | 2.10 | 0.05 | 1 | 0.24 | 3 |
| | OR | 0.66 | *** | *** | *** | *** | *** | 5.09 | 2.51 | 2.03 | 2 | -2.58 | *** |
| | RI | 0.09 | 1.50 | 2.30 | 0.65 | 1 | 1.37 | 0.94 | 2.22 | 0.42 | 1 | 5.61 | 2 |
| | UT | 1.15 | 0.80 | 2.02 | 0.40 | 1 | 2.03 | -3.15 | 2.10 | -1.50 | -2 | 0.03 | -1 |
| | VT | -0.27 | *** | *** | *** | *** | *** | 6.45 | 2.11 | 3.06 | 2 | 3.46 | *** |
| | WV | 0.90 | -2.33 | 2.06 | -1.13 | -2 | 4.07 | -4.64 | 1.95 | -2.39 | -2 | 2.74 | -4 |

21

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WI | -0.40 | 2.23 | 2.37 | 0.94 | 1 | 1.99 | *** | *** | *** | *** | *** | *** |
| | WY | -0.95 | -2.53 | 1.95 | -1.30 | -2 | -1.73 | 1.42 | 2.25 | 0.63 | 1 | -0.18 | -1 |

We begin by considering State Cohort Gains for the period 1992 to 1996. We note that the 15 high-stakes states with relevant data are not identical to the 15 high-stakes states in the previous section. Here, we have lost Ohio but gained Florida. We also now have data on 20 low-stakes states, rather than 18 earlier. We have lost Idaho and Oklahoma but gained Colorado, Delaware, Iowa, and Wisconsin. (Note: These "gains" and "losses" are entirely determined by the pattern of the states' participation in the NAEP assessments of 1992, 1996 and 2000.)

**Table 8a**
***Distribution of*** $\mathbf{W_1}$ ***for Hi-stakes and Lo-stakes States***

| # Lo-stakes states | $\mathbf{W_1}$ | # Hi-stakes states |
|---|---|---|
| 4 | **2** | 3 |
| 8 | **1** | 3 |
| 4 | **-1** | 2 |
| 4 | **-2** | 7 |

**Table 8b**
***Distribution of*** $\mathbf{W_2}$ ***for Hi-stakes and Lo-stakes States***

| # Lo-stakes states | $\mathbf{W_2}$ | # Hi-stakes states |
|---|---|---|
| 3 | **2** | 3 |
| 8 | **1** | 4 |
| 3 | **-1** | 4 |
| 5 | **-2** | 4 |

**Table 8c**
***Distribution of*** $\mathbf{W}$ ***for Hi-stakes and Lo-stakes States***

| # Lo-stakes states | **W** | # Hi-stakes states |
|---|---|---|
| 0 | **4** | 2 |
| 3 | **3** | 2 |
| 3 | **2** | 0 |
| 0 | **1** | 0 |
| 3 | **0** | 1 |
| 2 | **-1** | 4 |
| 1 | **-2** | 0 |
| 2 | **-3** | 3 |
| 2 | **-4** | 2 |

The distribution of $W_1$ for the high-stakes and low-stakes states is presented in Table 8a. The results for low-stakes states are somewhat better than those for high-stakes states. In particular, nearly half (7/15) of the latter experienced substantial negative gains ($W_1$ = -2). To develop further insight, we plot $g_1$ against c%ex (Figure 5) and against policy score (Figure 6). Neither plot contains any obvious outliers.

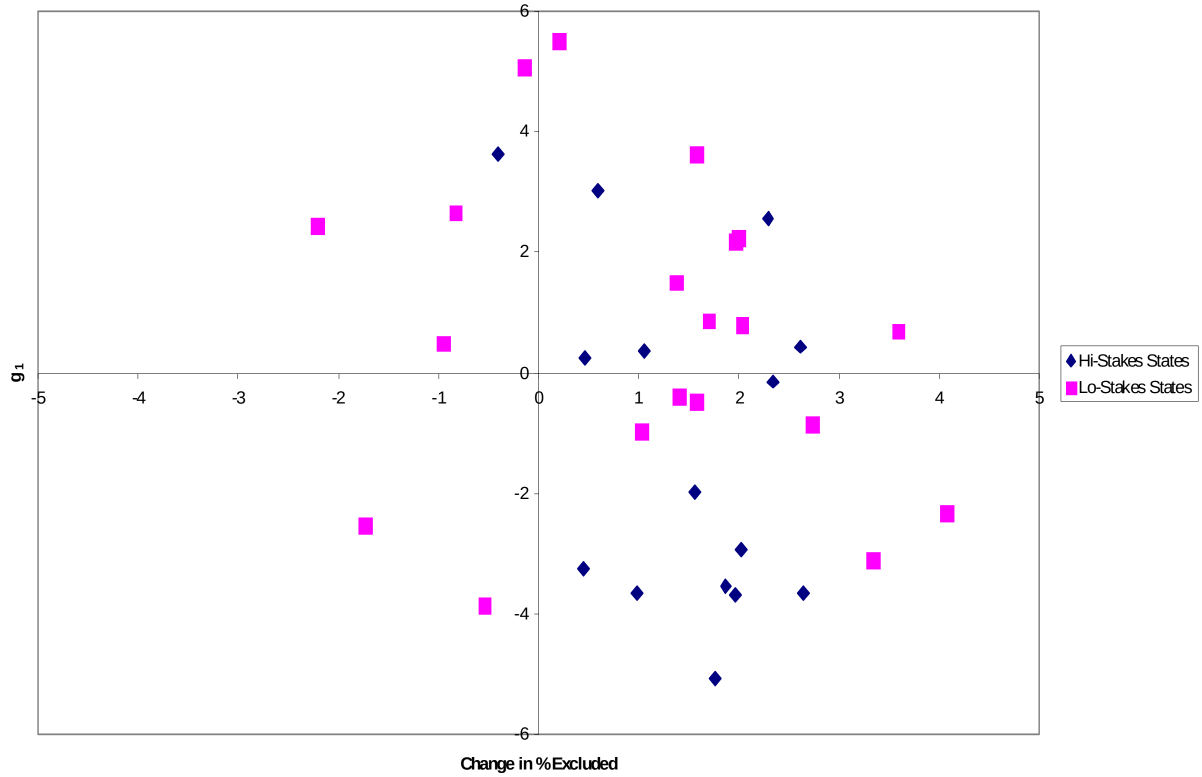


***Figure 5.*** **$g_1$ vs. Change in % excluded [Grade 4 (1992) to Grade 8 (1996)].**

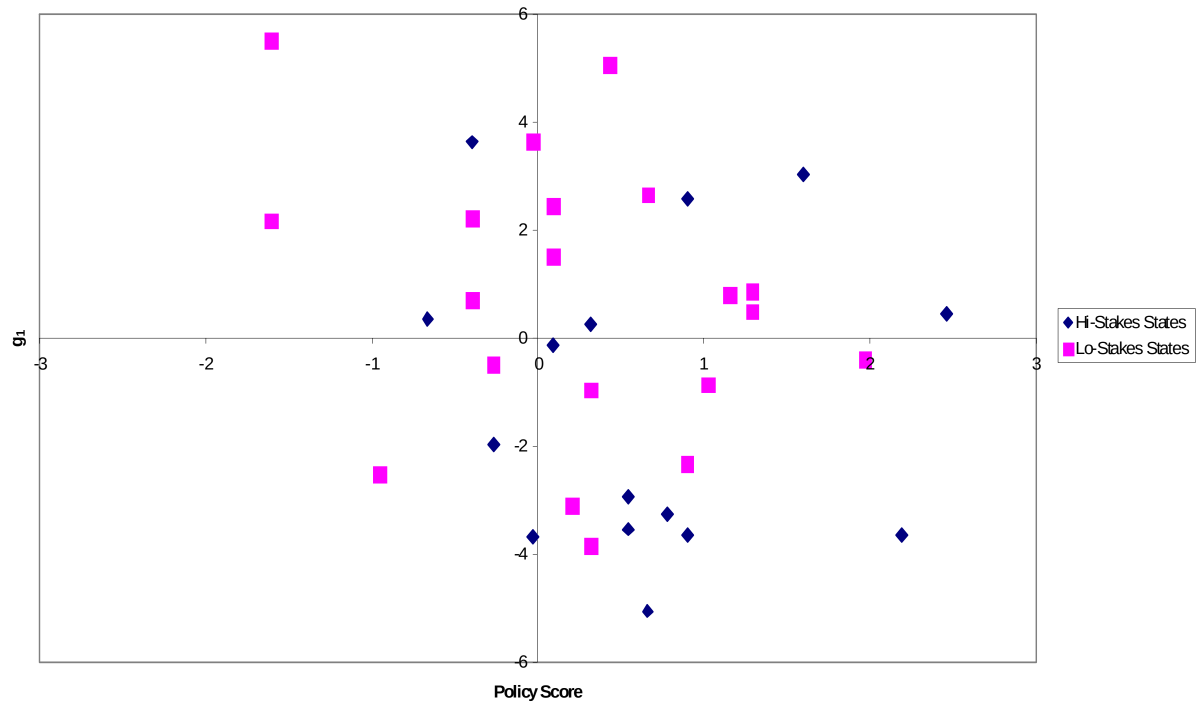


***Figure 6.*** $g_1$ **vs. Policy Score.**

Figure 5 suggests an overall negative association with c%ex and this is borne out by the correlations calculated separately for the two groups of states. (See Table 9.) Figure 6 also suggests an overall negative association with policy score and, again, this is borne out by the correlations for the two groups of states. (See Table 9.) Given the magnitude of the correlations, as well as the distributions of the two groups of states on the predictors, there does not appear to be an obvious "explanation" for the observed differences in results between the high-stakes and low-stakes states.

**Table 9**
***Correlations Between State Gains for 1992 Grade 4 to 1996 Grade 8 and Change in % Excluded and Policy Score***

| | Hi-stakes states | Lo-stakes states |
|---|---|---|
| Change in % excluded | -0.36 | -0.22 |
| Policy score | -0.06 | -0.28 |

Next we consider State Cohort Gains for the period 1996 to 2000. We note that again there are 15 high-stakes states with relevant data. In comparison to the previous section, we have lost Ohio but gained Nevada. We also now have data on 19 low-stakes states (rather than 18 earlier). We have lost Idaho and Oklahoma but gained Montana, Oregon, and Vermont.

The distribution of $W_2$ for high-stakes and low-stakes states is presented in Table 8b. The results for the low-stakes states are just slightly more positive than those for the high-stakes states. As before, we plot $g_2$ against c%ex (Figure 7) and against policy score (Figure 8). The plots for the later cohort are markedly different from those for the earlier cohort.

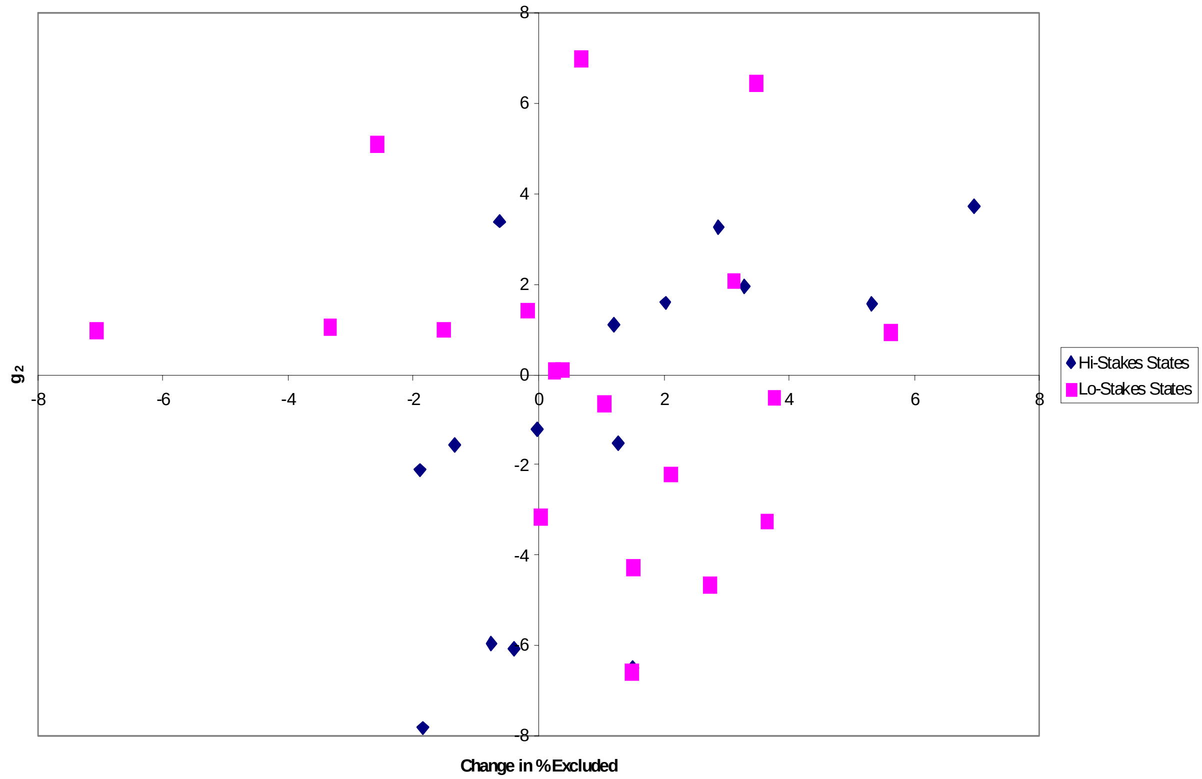


***Figure 7.*** **$g_2$ vs. Change in % Excluded [Grade 4 (1996) to Grade 8 (2000)].**

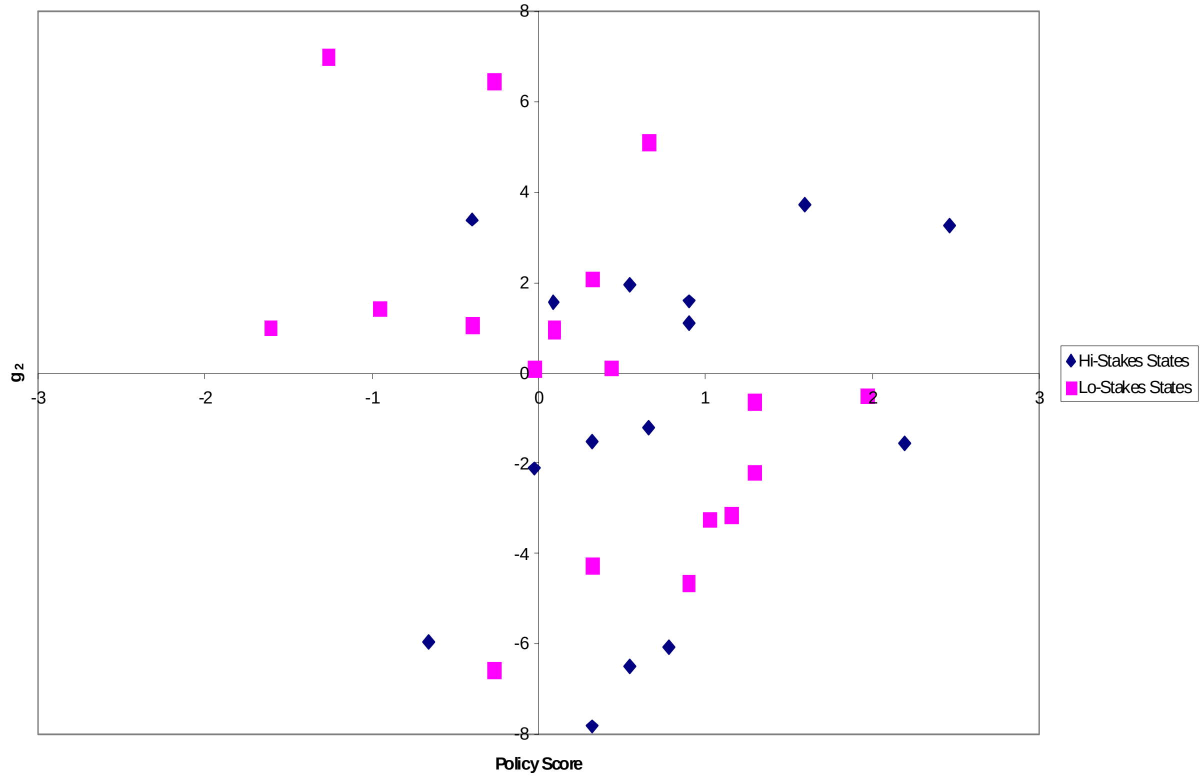


***Figure 8.*** **$g_2$ vs. Policy Score.**

In Figure 7, there is no clear overall pattern. When we calculate the correlations (see Table 10), we see a strong positive relationship for the high-stakes states and a weak negative relationship for the low-stakes states. Turning to Figure 8, we again see no strong overall pattern. However, the correlation with policy score is quite positive for the high-stakes states and quite negative for the low-stakes states (see Table 10).

**Table 10**
***Correlations Between State Gains for 1996 Grade 4 to 2000 Grade 8 and Change in % Excluded and Policy Score***

| | Hi-stakes states | Lo-stakes states |
|---|---|---|
| Change in % excluded | 0.61 | -0.18 |
| Policy score | 0.33 | -0.42 |

What are we to make of these results? With respect to the policy score, we would expect to see a stronger relationship to the gains for the later cohort, since the policy score is based on a review of state policies in 1996, the base year for the later cohort data. This is the case for the high-stakes states but not the low-stakes states. Indeed, the difference in the signs of the correlations for the two groups of states is puzzling and indicates some possible deficiencies in the formulation of the policy score. Specifically, the policy score does not have a strong accountability component.

The correlation of 0.61 in the second cohort between State Cohort Gain and change in percent excluded for high-stakes states is striking, and does not appear to be the result of a single outlier. This suggests that at least a portion of the gain in many of the high-stakes states may be attributable to increases in percent excluded. Making a regression adjustment (which implicitly maximizes the impact of the changes in percent excluded) would further increase the gap between high-stakes and low-stakes states, to the advantage of the latter. Finally, Table 8c, which displays the distribution of W for 14 high-stakes states and 16 low-stakes states with data for both cohorts, presents a decidedly mixed picture. Certainly, the lack of consistency in results between the cross-sectional and longitudinal analyses bears further scrutiny. The two most obvious problems are that the sets of states being compared differ somewhat and the criteria employed differ in both the grades and cohorts involved.

With regard to the first point, the exchanges of states between the two analyses appear to favor the low-stakes states rather than the high-stakes states. Consider, for example, the analysis of cohort gains for 1992 to 1996 (Table 8a). The high-stakes group lost Ohio, which was a high achieving state (V = 4) and gained Florida, which was a low achieving state ($W_1 = -1$). On the other hand, the low-stakes group lost Idaho and Oklahoma, which were both low achieving states (V = –4 for both) and gained Colorado, Delaware, Iowa, and Wisconsin, which were mostly high achieving states ( $W_1 = 2, -2, 1, 1$, respectively). A similar situation pertains to the cohort gains for 1996 to 2000 (Table 8b). The high-stakes group lost Ohio (V = 4) and gained Nevada ($W_2 = -1$). The low-stakes group lost Idaho and Oklahoma (V = –4 for both) and gained Montana, Oregon, and Vermont ($W_2 = 2$ for all).

Not surprisingly, if we examine the distributions of $W_1$ and $W_2$ for the same sets of high-stakes and low-stakes states used in the cross-sectional analysis, we find the comparisons do not favor one group over the other. It is tempting to go further and impute the missing $W_1$ and $W_2$ values (for the states in the cross-sectional analysis) based on their observed values of V. Were we to do so, we would find that the new distributions of $W_1$ are essentially identical and that the new distributions of $W_2$ slightly favor the high-stakes states. Such imputations, however, are themselves somewhat suspect since there is only a weak association between V and $W_1$ or between V and $W_2$. These findings are considered further in the Discussion.

**One Last Look**

Inasmuch as the intent of many reform efforts is to improve the achievement of low performing students, it seems worthwhile to investigate the patterns of relative improvement at a point other than the means of the score distributions. Accordingly, we selected the 25$^{th}$ percentile and carried through analyses that parallel those reported above. The basic data are presented in Table A3 of the appendix.

First, for both grades 4 and 8, we compute the state gain at the 25$^{th}$ percentile compared to the nation, over the period 1992 to 2000. We normalize the state gains by dividing by the estimated standard errors. For grade 4, we denote the derived statistic by $D'_4$ and for grade 8 by $D'_8$. These statistics are plotted against $D_4$ and $D_8$, respectively, in Figures 9 and 10. As before, we distinguish high-stakes and low-stakes states and now fit separate least squares lines to the data in each figure. We can see that the comparisons between the two groups of states based on results at the 25$^{th}$ percentiles are very similar to those based on the means.

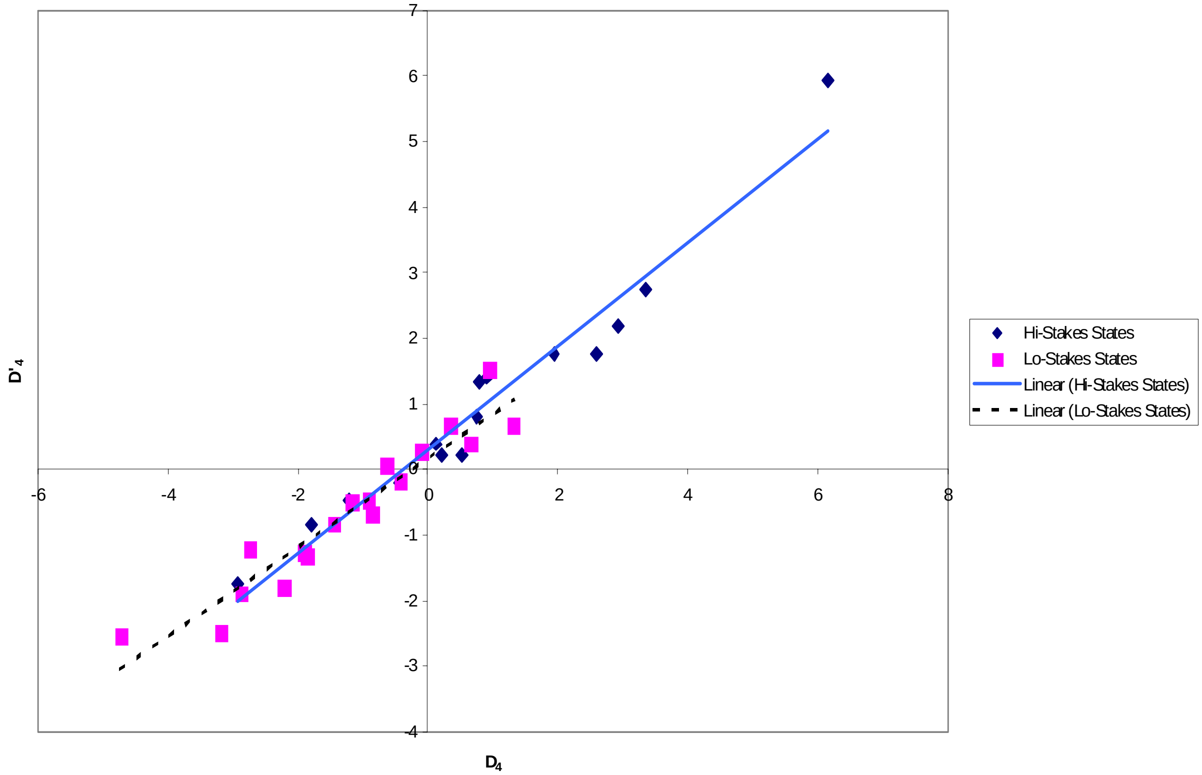


***Figure 9.*** **Grade 4: $D_4'$ vs. $D_4$.**

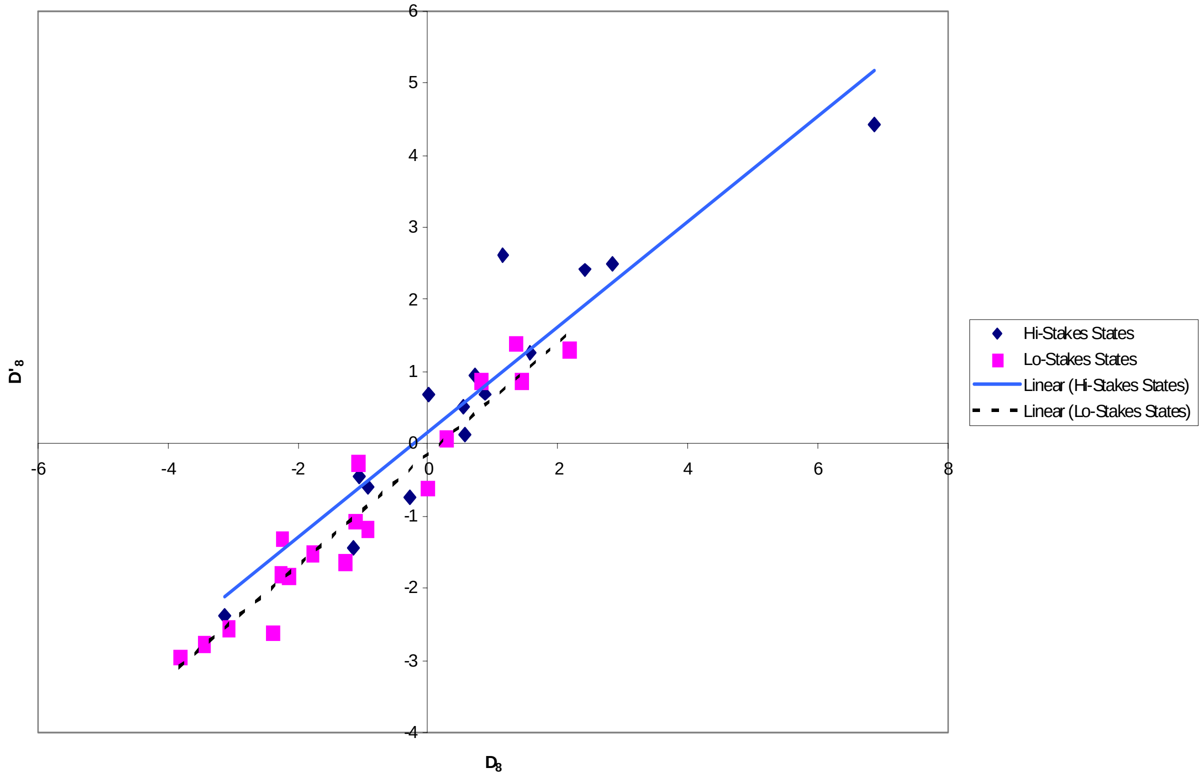


***Figure 10.*** **Grade 8:** $D'_8$ **vs.** $D_8$**.**

Second, we carried out a cohort analysis using the 25th percentile. The basic data are presented in Table A4 of the appendix. We computed the state cohort gain (grade 4, 1992 to grade 8, 1996) at the 25th percentile compared to the nation. After dividing by the estimated standard error, we obtained a statistic, which we denote by $G'_1$. For the state cohort gain (grade 4, 1996 to grade 8, 2000), the corresponding statistic is denoted by $G'_2$. These statistics are plotted against $G_1$ and $G_2$, respectively in Figures 11 and 12. Again, we distinguish high-stakes and low-stakes states and fit separate least squares lines to the data in each figure. The figures indicate that the cohort results at the 25th percentile are similar to those at the mean.

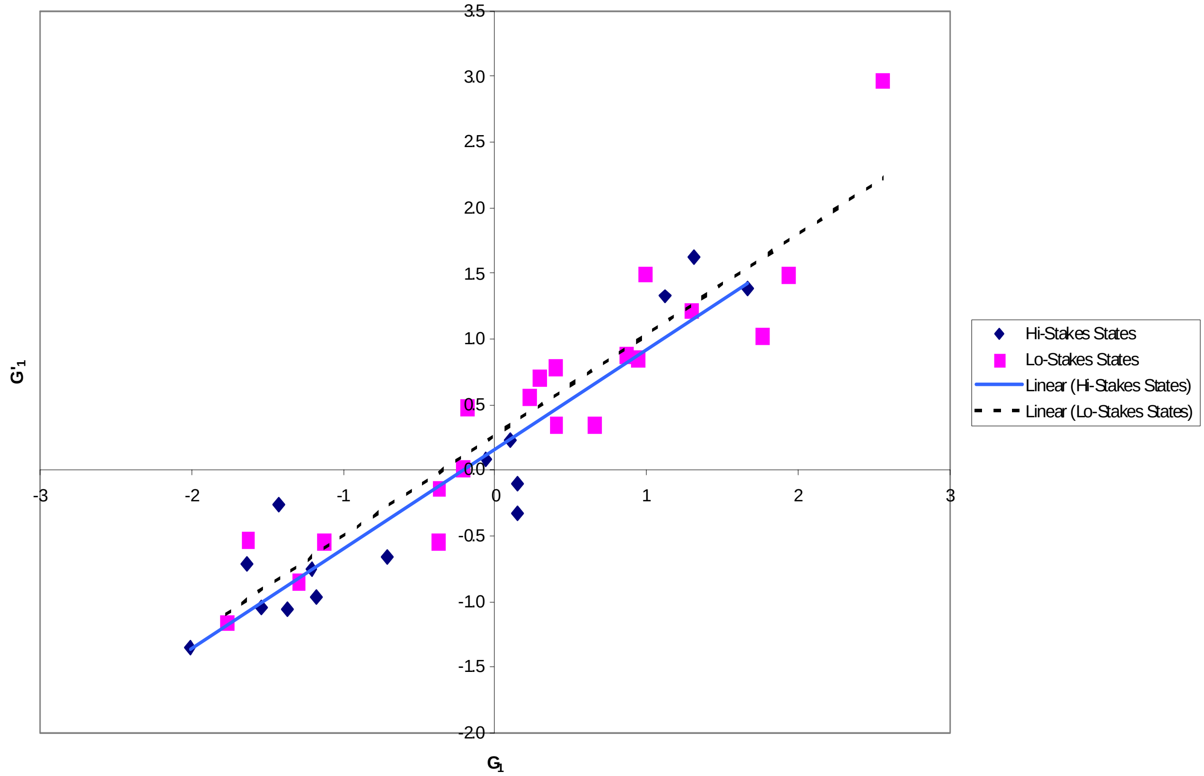


***Figure 11.*** **$G_1'$ vs. $G_1$ [Grade 4 (1992) to Grade 8 (1996)].**

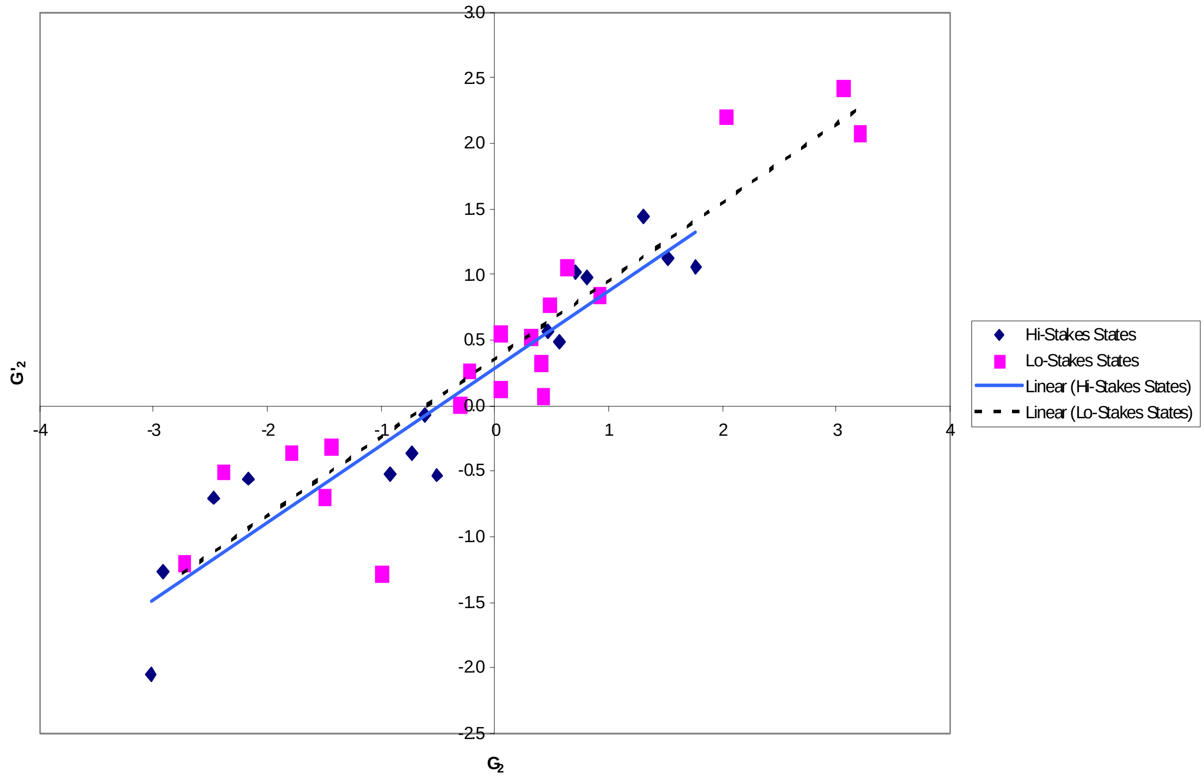


***Figure 12.*** $G_2'$ **vs.** $G_2$ **[Grade 4 (1996) to Grade 8 (2000)].**

**Discussion**

Our extended reanalysis of the NAEP component of Amrein and Berliner (2002a) began with a comparison of high-stakes and low-stakes states in terms of their relative gains over time at grades four and eight. These repeated cross-sectional contrasts strongly favor the high-stakes states, both when we look at the raw relative gains at the mean and when we consider the standard errors associated with those gains. The results cannot be accounted for by differences between the two groups in changes in percent excluded over that period or by the correlation between the relative gains and the change in percent excluded. There is some robustness to our findings inasmuch as the analyses at the 25$^{th}$ percentile produced similar results. Moreover, consideration of a state reform policy score due to Swanson and Stevenson also fails to account for differences in relative gains between states. Consequently, our conclusions differ from those in Amrein and Berliner (2002a).

Other investigators have also taken issue with Amrein and Berliner. Carnoy and Loeb (2003), for example, adopt a somewhat different methodological approach. They focus on the period 1996 to 2000 and consider a number of different education outcomes. With respect to NAEP math scores, the criteria are the proportions of students, in grades 4 and 8, meeting the basic or proficient standards. They develop an accountability index, with each state assigned a score from 0 to 5 based on the estimated strength of its accountability efforts. (Note: This differs from the policy score of Swanson and Stevenson, which is a broader measure of a state's education reform efforts.)

They do not divide the states into two groups; rather, they carry out a regression with the state as the unit of analysis. The criterion is the change in percent above the standard over the period 1996 to 2000 and the predictors are the accountability index for the state, as well as a number of other political, demographic and educational variables. The latter variables are selected on the basis of a preliminary regression in which the accountability index serves as the criterion and a large set of state characteristics are potential predictors.

The rationale for this approach is quite sensible. To quote Carnoy and Loeb (p. 5): "… variables that could influence both the strength of accountability reforms and student outcomes are relevant." That is, including such variables, along with the accountability index, in the main regression should reduce the misspecification bias and lead to more accurate estimates of the strength of association between the change in percent above the standard and the accountability index. Moreover, they estimate these regressions separately for White, Black, and Hispanic students.

They find a relatively strong positive association between gains and the accountability index in grade 8, especially for Black and Hispanic students but a much weaker, though still positive, association in grade 4 (except for Black students, where the association is as strong as in grade 8). Consider, for example, the results for the 8$^{th}$ grade. On average, states that differ by two points on the accountability index (controlling for the other variables) exhibit non-trivial differences in gains (over the period 1996 to 2000) in the percentages of students achieving the basis skills level: For White students it is 2.8 percentage points; for Black students it is 5.1 percentage points; for Hispanic students it is 8.9 percentage points. Actually, these findings are consistent with the results of Amrein and Berliner (2002a) for the period 1996 – 2000. The latter found a much stronger relationship between testing policy and relative score gains in the 8$^{th}$ grade than in the 4$^{th}$ grade.

Since our analysis of relative gains at the mean focused on changes from 1992 to 2000, we replicated it for the period 1996 to 2000 to provide a more direct comparison to Carnoy and Loeb. We do not report the results here but note that we again observed a substantial advantage for high-stakes states, equally strong in both grades 4 and 8. When we plot the analogs of $D_4$ and $D_8$ against the policy score of Swanson and Stevenson, we find essentially no relationship in grade 4 and a moderately positive one in grade 8.

Despite substantial methodological differences in the two approaches, the general tenor of the findings in Carnoy and Loeb with respect to NAEP results is consistent with ours: With the data available, there is no basis for rejecting the inference that the introduction of high-stakes testing for accountability is associated with gains in NAEP mathematics achievement through the 1990s. Moreover, the strength of the association between states' gains and a measure of the general accountability efforts in the states is greater in the 8$^{th}$ grade than in the 4$^{th}$.

In a subsequent paper, Amrein and Berliner (2002b), present an extensive compilation of the history of policy efforts in each state. Employing what they term an archival time series research design, Amrein and Berliner investigate the relationship between those efforts and subsequent relative gains (or losses) in NAEP achievement. In this context, as many as 28 states are identified as high-stakes testing states, where high-stakes is taken to mean that consequences have been attached to test results beyond those in place for many years.

Their analysis, however, is undermined by a too-zealous use of changes in exclusion rates as a basis for eliminating states from consideration. For example, differential gains by North

Carolina are essentially dismissed because exclusion rates also increased over the period in question. This is done without regard to the relative (to other states) magnitudes of these changes. The number of states remaining is rather small and, thus, their conclusion that "…there is inadequate evidence to support the proposition that high-stakes tests and high school graduation exams increase student achievement" (p. 57) appears unwarranted. Raymond and Hanushek (2003) are very critical of Amrein and Berliner (2002b) on these and other grounds. Interested readers can find a related exchange of letters in the Fall 2003 issue of Education Next.

(Note: After the original version of this report was prepared, the author learned of a paper by Rosenshine (2003). He also takes issue with the methodology of Amrein and Berliner (2000b) and carries out an alternative analysis comparing high-stakes and low-stakes states. For the period 1996 to 2000, he finds greater improvement for the high-stakes states in 4th grade and 8th grade mathematics. In response, Amrein-Beardsley and Berliner (2003) conducted a reanalysis, with similar conclusions. However, they discount the findings because of the tendency of high-stakes states to have greater increases in the percent of students excluded from the NAEP assessments.)

As we acknowledged at the outset, conclusions from analyses concerning the effects of testing or, more generally, accountability policies must be tentative, based as they are on highly aggregated, observational data. This caution proves well founded for the picture is somewhat different when we turn from cross-sectional analyses to pseudo-longitudinal studies, which involve following two different cohorts from grade 4 to grade 8. For both cohorts, the comparisons of relative gains slightly favor low-stakes states. Moreover, only part of the shift in favor of low-stakes states can be explained by the changes in the sets of states contributing to the different analyses. We also note that for the later cohort, there is a strong positive correlation between relative gains and change in percent excluded for the high-stakes states but a strong negative correlation for the low-stakes states. Our analysis of the cohort data leads to a conclusion (again tentative) that is consistent with that of Amrein and Berliner (2002a).

Raymond and Hanushek (2003) also carry out a cohort analysis, but find a slight advantage in favor of states with higher stakes attached to test results. How can we explain these differences? First, with regard to the cohort analyses, Raymond and Hanushek are neither explicit about the criteria they employ to categorize states nor do they present the numbers of states in each category. It is likely that their three categories do not mesh cleanly with the two sets of 18 and 32 states that were the basis our cohort analysis. Equally important, Raymond and Hanushek indicate that they adjusted their results "…to account for changes in state spending on education and for parents' education levels…" (p. 54). They do not describe the impact of these adjustments on the results.

Presumably, these differences in findings for the cohorts can be explained once the categorization of states and the factors used in adjustment are elucidated. It is not obvious, however, how the adjusted results should be interpreted when viewed in the context of the changes in exclusion rates observed for each category of states. We must also remember that test scores are only one aspect of an educational system's output and that other measures should be accorded comparable attention.

The apparent inconsistency between the cross-sectional and longitudinal approaches is a signal for the larger question of which is to be preferred for policy purposes. The former is more common and is the basis of the adequate yearly progress provisions of NCLB. The latter is more appealing (at least to some) in that it avoids certain confounds and appears to be a fairer and more meaningful measure of the contribution of the system to student achievement. It also is gaining popularity among state education departments. If one accepts cross-sectional analyses as the coin of the realm, then it appears that high-stakes testing is strongly associated with larger gains over the period 1992 to 2000. Those who believe that this does not imply a causal link must offer (perhaps) a different criterion, an alternative categorization of states, additional explanatory factors, a new methodology—or some combination of the foregoing. Apparently, this has not been done.

Of course, there are other approaches. For example, one could try to show that the costs and consequences related to high-stakes testing (expenditures, student dropouts, teacher attrition, etc.) are not worth the score gains, particularly if those gains are not strongly tied to valued learning goals. Carrying out such a cost- benefit analysis would be a challenging path to take (Levin & McEwan, 2000). It would be worthwhile, however, as it might provide policymakers with a broader foundation for considering policy alternatives.

On the other hand, if the logic of pseudo-longitudinal studies is more persuasive, then the argument in favor of high-stakes testing is more difficult to make and the burden of "proof" now falls on its proponents—who also have a variety of methodological options to choose from. In either case, as one reviewer has remarked, it is both disappointing and troubling that the trajectories at the $25^{th}$ percentiles so closely track the trajectories at the mean. Many reform initiatives, most prominently Title I, have been directed at lower performing students – yet these students do not appear to have derived any special benefit.

The general lack of strong associations across states between achievement gains and policy scores suggests that we have to be more diligent in documenting each state's policy history as well as the trajectories through time of other relevant variables. Perhaps what is required is a multidimensional representation of the states' education policies that takes explicit account of the time sequence of the various initiatives as well as the scope and quality of implementation. We would then have to appropriately incorporate that information into our analyses, exercising due respect for the limitations of both the design and the data.

Nonetheless, no matter how careful and comprehensive we are, it is certain that the process of drawing defensible policy conclusions in this context will be fraught with difficulty and controversy. Even apparently sensible advice about educational reform (e.g. Barton, 2002) should be put to empirical test. However, given a limited observational database to work with, there are many policy indices and numerous ancillary variables that can be used in different ways to support or debunk the efficacy of high-stakes testing efforts or any other reform initiative. If we acknowledge that most states have embarked on a number of initiatives that, to a greater or lesser extent, overlap in time, then we must recognize that attributing observed differences in results to one of those initiatives is very problematic.

In a similar vein, another reviewer pointed out, commenting on Table 8c for the cohort analysis, that states could have been classified according to whether or not they had been members of the Confederate States of America (CSA). Ninety percent of the CSA members had negative gains but only about a third of the other states did. Clearly, this is not to be taken seriously and no doubt there are other bases for classifying states (more or less plausible) that yield even more striking contrasts. The point, again, is to illustrate the inferential problems we face when it is possible to generate many more hypotheses than the data can properly address. Consequently, the stories we glean from large-scale education databases of this sort will likely be a great deal more Delphic than either researchers or policymakers would prefer.

**Acknowledgements**

The author would to thank John Willey for generating the tables and graphs and Elizabeth Brophy and William Monaghan for preparing the final draft of the paper. He also appreciates the useful comments and suggestions received from Brent Bridgeman, Dan Eignor, and Richard Phelps on an earlier version.

## References

Amrein, A. L., & Berliner, D. C. (2002a). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved March 28, 2002, from http://epaa.asa.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*. Retrieved December 2002, from Educational Policy Studies Laboratory, Education Policy Research Unit: http://edpolicylab.org

Amrein-Beardsley, A., & Berliner, D. C. (2003, August). 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 September 21, 2003, from http://epaa.asu.edu/epaa/v11n25/

Barton, P. (2002). *Staying on course in education reform* (ETS Policy Information Center Perspective). Princeton, NJ: Educational Testing Service.

Barton, P., & Coley, R. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade* (ETS Policy Information Center Report). Princeton, NJ: Educational Testing Service.

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Elmore, R. (2002, Spring). Unwarranted intrusion. *Education Next, 2*(1), 31-35. Retrieved March 14, 2003, from http://www.educationnext.org/20021/30.html

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP scores tell us*. Santa Monica, CA: RAND Corporation.

Levin, H. M., & McEwan, P. J. (Eds.) (2000). *Cost-effectiveness analysis: Methods and applications* (2nd ed.).Thousand Oaks, CA: Sage.

Lewis, A. C. (2002, November). A horse called NCLB. *Phi Delta Kappan*, 179-180.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher 29*(2), 4-16.

Mehrens, W. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13). Retrieved September 3, 2002, from http://epaa.asu.edu/epaa/v6nl3

National Center for Education Statistics. (2003). *National Assessment of Educational Progress—The Nation's Report Card.* Retrieved October 17, 2003, from http://nces.ed.gov/nationsreportcard/

Raudenbush, S. W., Fotiu, R. P., Cheong, Y. F., & Ziazi, Z. M. (1995, July). *Synthesizing results from the Trial State Assessment*. Paper presented at the Joint Statistical Meetings.

Raymond, M .E., & Hanushek, E. A (2003, Summer). High-stakes research. *Education Next, 3*(3), 48-55. Retrieved September 21, 2003, from http://www.educationnext.org

Rosenshine, B. (2003, August). High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved August 21, 2003, from http://epaa.asu.edu.epaa/v11n24/

Steinberg, L. (2003, February 5). Does high-stakes testing hurt students? *Education Week*, p. 48.

Swanson, C. B., & Stevenson, D. L. (2002, Spring). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis*, *24*(1), pp. 1-27.

Wolf, D. P. (2002, December). When raising isn't rising: The failure of accountability systems to measure student growth over time. *American Association of School Administrations—The School Administrator Web Edition*. Retrieved January 20, 2003, from http://www.aasa.org/publications/sa/2002_12/Wolf.htm

# Appendix

## Table A1
***Overall Means for NAEP Data for Grades 4 and 8***

| | | | Grade 4 | | | | | | Grade 8 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1992 | | | 2000 | | | 1992 | | | 2000 | | |
| | State | Abbrev. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. |
| **Hi-stakes states** | **National (Public)** | **NAT** | **218.58** | **0.80** | **6.52** | **226.24** | **0.96** | **7.47** | **266.87** | **0.96** | **6.39** | **274.42** | **0.84** | **7.39** |
| | **Alabama** | **AL** | **208.33** | **1.56** | **4.51** | **217.94** | **1.41** | **5.78** | **252.19** | **1.66** | **5.45** | **262.16** | **1.77** | **4.94** |
| | **Georgia** | **GA** | **215.59** | **1.23** | **5.23** | **219.56** | **1.06** | **6.51** | **259.36** | **1.16** | **4.68** | **266.33** | **1.25** | **7.20** |
| | **Indiana** | **IN** | **221.04** | **1.04** | **3.32** | **234.42** | **1.08** | **6.78** | **270.10** | **1.14** | **4.58** | **283.05** | **1.45** | **7.30** |
| | **Louisiana** | **LA** | **204.14** | **1.46** | **4.00** | **217.96** | **1.40** | **7.68** | **249.98** | **1.66** | **4.27** | **258.98** | **1.50** | **5.76** |
| | **Maryland** | **MD** | **217.32** | **1.28** | **4.05** | **222.31** | **1.27** | **8.93** | **264.83** | **1.28** | **4.68** | **276.01** | **1.43** | **10.56** |
| | **Minnesota** | **MN** | **228.49** | **0.90** | **3.36** | **235.27** | **1.32** | **5.80** | **282.39** | **0.96** | **3.46** | **287.65** | **1.44** | **5.49** |
| | **Mississippi** | **MS** | **201.83** | **1.08** | **4.85** | **210.97** | **1.07** | **4.26** | **246.46** | **1.18** | **7.17** | **254.03** | **1.30** | **7.47** |
| | **New Mexico** | **NM** | **213.30** | **1.44** | **7.35** | **213.87** | **1.48** | **12.30** | **259.61** | **0.90** | **5.34** | **259.84** | **1.74** | **11.55** |
| | **New York** | **NY** | **218.45** | **1.25** | **5.29** | **226.56** | **1.33** | **11.53** | **266.42** | **2.08** | **8.49** | **276.26** | **2.09** | **13.13** |
| | **North Carolina** | **NC** | **212.88** | **1.09** | **3.86** | **232.46** | **1.00** | **13.34** | **258.41** | **1.17** | **3.26** | **280.13** | **1.13** | **13.85** |
| | **Ohio** | **OH** | **218.71** | **1.18** | **6.04** | **230.57** | **1.33** | **9.95** | **268.12** | **1.52** | **6.09** | **282.67** | **1.48** | **8.70** |
| | **South Carolina** | **SC** | **212.50** | **1.08** | **4.81** | **220.42** | **1.39** | **7.46** | **260.77** | **0.97** | **5.91** | **266.35** | **1.39** | **6.85** |
| | **Tennessee** | **TN** | **210.95** | **1.35** | **3.87** | **219.84** | **1.49** | **3.77** | **258.83** | **1.39** | **4.89** | **263.44** | **1.72** | **4.56** |
| | **Texas** | **TX** | **217.92** | **1.21** | **7.58** | **232.67** | **1.21** | **15.44** | **264.59** | **1.30** | **6.58** | **274.85** | **1.47** | **9.57** |
| | **Virginia** | **VA** | **220.76** | **1.30** | **5.24** | **230.39** | **1.27** | **10.79** | **267.86** | **1.16** | **5.28** | **276.67** | **1.50** | **9.97** |
| Lo-stakes states | National (Public) | NAT | 218.58 | 0.80 | 6.52 | 226.24 | 0.96 | 7.47 | 266.87 | 0.96 | 6.39 | 274.42 | 0.84 | 7.39 |
| | Arizona | AZ | 215.25 | 1.07 | 4.93 | 218.77 | 1.42 | 11.85 | 265.37 | 1.26 | 5.71 | 270.72 | 1.53 | 9.08 |
| | Arkansas | AR | 210.21 | 0.89 | 5.43 | 217.06 | 1.13 | 6.58 | 256.31 | 1.19 | 6.24 | 261.36 | 1.37 | 8.18 |
| | California | CA | 208.40 | 1.56 | 12.20 | 213.57 | 1.84 | 8.96 | 260.89 | 1.66 | 8.26 | 262.17 | 2.04 | 8.68 |
| | Connecticut | CT | 226.80 | 1.13 | 6.58 | 234.24 | 1.16 | 9.95 | 273.74 | 1.14 | 6.59 | 281.90 | 1.37 | 10.19 |
| | Hawaii | HI | 214.06 | 1.31 | 5.75 | 215.85 | 1.15 | 10.17 | 257.41 | 0.86 | 4.86 | 262.77 | 1.34 | 7.28 |
| | Idaho | ID | 221.56 | 0.96 | 3.45 | 226.89 | 1.21 | 5.91 | 275.09 | 0.74 | 3.17 | 277.92 | 1.31 | 4.76 |
| | Kentucky | KY | 215.05 | 1.01 | 3.29 | 220.99 | 1.17 | 8.24 | 262.24 | 1.11 | 4.56 | 271.56 | 1.40 | 9.47 |
| | Maine | ME | 231.64 | 1.00 | 5.66 | 230.57 | 0.92 | 10.12 | 278.64 | 0.98 | 4.43 | 283.64 | 1.19 | 8.56 |
| | Massachusetts | MA | 226.60 | 1.17 | 6.99 | 234.96 | 1.12 | 10.44 | 272.78 | 1.05 | 8.14 | 283.12 | 1.25 | 12.13 |
| | Michigan | MI | 219.88 | 1.71 | 5.26 | 230.89 | 1.43 | 8.37 | 267.35 | 1.39 | 6.06 | 278.45 | 1.60 | 6.54 |
| | Missouri | MO | 222.22 | 1.19 | 4.37 | 228.55 | 1.19 | 9.71 | 271.13 | 1.19 | 4.47 | 273.58 | 1.46 | 8.58 |
| | Nebraska | NE | 225.33 | 1.23 | 4.13 | 225.95 | 1.72 | 7.58 | 277.65 | 1.11 | 4.07 | 280.62 | 1.12 | 3.50 |
| | North Dakota | ND | 228.66 | 0.77 | 1.74 | 230.89 | 0.86 | 5.72 | 283.21 | 1.14 | 2.40 | 283.07 | 1.07 | 3.80 |
| | Oklahoma | OK | 220.32 | 0.98 | 7.29 | 225.04 | 1.26 | 10.47 | 268.13 | 1.15 | 6.30 | 271.65 | 1.48 | 8.61 |
| | Rhode Island | RI | 215.45 | 1.53 | 5.81 | 224.63 | 1.22 | 11.70 | 265.91 | 0.73 | 5.10 | 273.43 | 1.11 | 11.77 |
| | Utah | UT | 224.04 | 0.97 | 4.03 | 227.29 | 1.22 | 6.66 | 274.34 | 0.73 | 4.37 | 275.44 | 1.16 | 5.82 |
| | West Virginia | WV | 215.27 | 1.05 | 4.42 | 224.85 | 1.20 | 10.07 | 259.09 | 1.01 | 5.79 | 270.78 | 1.00 | 11.08 |
| | Wyoming | WY | 225.38 | 0.93 | 3.52 | 229.25 | 1.30 | 6.07 | 275.08 | 0.86 | 3.95 | 276.69 | 1.18 | 3.96 |

## Table A2
### *Overall Means for NAEP Data for Cohort Analysis*

| | | | Cohort (1992 to 1996) | | | | | | Cohort (1996 to 2000) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Grade 4 (1992) | | | Grade 8 (1996) | | | Grade 4 (1996) | | | Grade 8 (2000) | | |
| | State | Abbrev. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. |
| Hi-stakes states | **National (Public)** | **NAT** | **218.58** | **0.80** | **6.52** | **270.51** | **1.21** | **4.84** | **222.34** | **1.03** | **6.36** | **274.42** | **0.84** | **7.39** |
| | **Alabama** | **AL** | **208.33** | **1.56** | **4.51** | **256.59** | **2.15** | **7.15** | **211.65** | **1.24** | **6.28** | **262.16** | **1.77** | **4.94** |
| | **Florida** | **FL** | **213.69** | **1.50** | **8.26** | **263.64** | **1.84** | **9.82** | **215.76** | **1.16** | **9.95** | *** | *** | *** |
| | **Georgia** | **GA** | **215.59** | **1.23** | **5.23** | **262.47** | **1.65** | **6.99** | **215.46** | **1.49** | **7.23** | **266.33** | **1.25** | **7.20** |
| | **Indiana** | **IN** | **221.04** | **1.04** | **3.32** | **275.53** | **1.44** | **5.61** | **229.39** | **1.05** | **5.28** | **283.05** | **1.45** | **7.30** |
| | **Louisiana** | **LA** | **204.14** | **1.46** | **4.00** | **252.38** | **1.57** | **5.96** | **209.02** | **1.11** | **7.66** | **258.98** | **1.50** | **5.76** |
| | **Maryland** | **MD** | **217.32** | **1.28** | **4.05** | **269.68** | **2.13** | **6.67** | **220.69** | **1.56** | **7.68** | **276.01** | **1.43** | **10.56** |
| | **Minnesota** | **MN** | **228.49** | **0.90** | **3.36** | **284.05** | **1.34** | **2.96** | **232.19** | **1.08** | **6.12** | **287.65** | **1.44** | **5.49** |
| | **Mississippi** | **MS** | **201.83** | **1.08** | **4.85** | **250.22** | **1.19** | **6.72** | **208.43** | **1.22** | **5.97** | **254.03** | **1.30** | **7.47** |
| | **Nevada** | **NV** | *** | *** | *** | *** | *** | *** | **217.62** | **1.30** | **8.90** | **268.18** | **0.94** | **10.17** |
| | **New Mexico** | **NM** | **213.30** | **1.44** | **7.35** | **261.97** | **1.22** | **7.80** | **213.84** | **1.75** | **11.94** | **259.84** | **1.74** | **11.55** |
| | **New York** | **NY** | **218.45** | **1.25** | **5.29** | **270.23** | **1.66** | **7.63** | **222.63** | **1.24** | **7.80** | **276.26** | **2.09** | **13.13** |
| | **North Carolina** | **NC** | **212.88** | **1.09** | **3.86** | **267.83** | **1.42** | **4.45** | **224.33** | **1.19** | **6.89** | **280.13** | **1.13** | **13.85** |
| | **South Carolina** | **SC** | **212.50** | **1.08** | **4.81** | **260.78** | **1.54** | **5.79** | **213.19** | **1.30** | **5.65** | **266.35** | **1.39** | **6.85** |
| | **Tennessee** | **TN** | **210.95** | **1.35** | **3.87** | **263.12** | **1.40** | **4.33** | **219.18** | **1.40** | **6.40** | **263.44** | **1.72** | **4.56** |
| | **Texas** | **TX** | **217.92** | **1.21** | **7.58** | **270.20** | **1.43** | **8.64** | **228.71** | **1.36** | **10.32** | **274.85** | **1.47** | **9.57** |
| | **Virginia** | **VA** | **220.76** | **1.30** | **5.24** | **269.75** | **1.56** | **7.26** | **222.64** | **1.36** | **6.68** | **276.67** | **1.50** | **9.97** |
| Lo-stakes states | National (Public) | NAT | 218.58 | 0.80 | 6.52 | 270.51 | 1.21 | 4.84 | 222.34 | 1.03 | 6.36 | 274.42 | 0.84 | 7.39 |
| | Arizona | AZ | 215.25 | 1.07 | 4.93 | 267.87 | 1.56 | 8.52 | 217.58 | 1.73 | 12.42 | 270.72 | 1.53 | 9.08 |
| | Arkansas | AR | 210.21 | 0.89 | 5.43 | 261.65 | 1.52 | 7.01 | 215.85 | 1.46 | 6.69 | 261.36 | 1.37 | 8.18 |
| | California | CA | 208.40 | 1.56 | 12.20 | 262.77 | 1.85 | 9.99 | 209.13 | 1.84 | 15.75 | 262.17 | 2.04 | 8.68 |
| | Colorado | CO | 221.02 | 0.97 | 5.24 | 275.61 | 1.09 | 4.41 | 225.81 | 1.04 | 8.38 | *** | *** | *** |
| | Connecticut | CT | 226.80 | 1.13 | 6.58 | 279.59 | 1.12 | 8.28 | 232.03 | 1.10 | 8.09 | 281.90 | 1.37 | 10.19 |
| | Delaware | DE | 217.90 | 0.78 | 5.30 | 266.73 | 0.95 | 8.64 | 215.03 | 0.64 | 7.04 | *** | *** | *** |
| | Hawaii | HI | 214.06 | 1.31 | 5.75 | 262.13 | 0.97 | 5.21 | 214.97 | 1.45 | 5.78 | 262.77 | 1.34 | 7.28 |
| | Iowa | IA | 229.88 | 1.02 | 3.26 | 283.99 | 1.31 | 5.23 | 229.13 | 1.08 | 5.62 | *** | *** | *** |
| | Kentucky | KY | 215.05 | 1.01 | 3.29 | 266.59 | 1.07 | 4.69 | 219.99 | 1.07 | 5.72 | 271.56 | 1.40 | 9.47 |
| | Maine | ME | 231.64 | 1.00 | 5.66 | 284.06 | 1.29 | 4.71 | 232.21 | 1.02 | 7.52 | 283.64 | 1.19 | 8.56 |
| | Massachusetts | MA | 226.60 | 1.17 | 6.99 | 277.57 | 1.74 | 8.02 | 228.97 | 1.35 | 9.02 | 283.12 | 1.25 | 12.13 |
| | Michigan | MI | 219.88 | 1.71 | 5.26 | 276.87 | 1.79 | 5.12 | 226.26 | 1.27 | 6.17 | 278.45 | 1.60 | 6.54 |
| | Missouri | MO | 222.22 | 1.19 | 4.37 | 273.28 | 1.39 | 7.10 | 224.73 | 1.07 | 4.94 | 273.58 | 1.46 | 8.58 |
| | Montana | MT | *** | *** | *** | 283.00 | 1.30 | 3.22 | 227.52 | 1.23 | 4.77 | 286.58 | 1.22 | 5.44 |
| | Nebraska | NE | 225.33 | 1.23 | 4.13 | 282.77 | 1.02 | 4.33 | 227.54 | 1.18 | 5.02 | 280.62 | 1.12 | 3.50 |
| | North Dakota | ND | 228.66 | 0.77 | 1.74 | 284.22 | 0.91 | 3.32 | 230.90 | 1.23 | 3.56 | 283.07 | 1.07 | 3.80 |
| | Oregon | OR | *** | *** | *** | 276.34 | 1.47 | 3.92 | 223.48 | 1.35 | 8.77 | 280.64 | 1.65 | 6.19 |
| | Rhode Island | RI | 215.45 | 1.53 | 5.81 | 268.88 | 0.92 | 7.18 | 220.42 | 1.39 | 6.16 | 273.43 | 1.11 | 11.77 |
| | Utah | UT | 224.04 | 0.97 | 4.03 | 276.77 | 1.03 | 6.06 | 226.52 | 1.15 | 5.79 | 275.44 | 1.16 | 5.82 |
| | Vermont | VT | *** | *** | *** | 279.25 | 0.95 | 4.29 | 224.88 | 1.22 | 6.08 | 283.41 | 1.10 | 9.54 |
| | West Virginia | WV | 215.27 | 1.05 | 4.42 | 264.87 | 1.02 | 8.49 | 223.35 | 1.01 | 8.34 | 270.78 | 1.00 | 11.08 |
| | Wisconsin | WI | 228.69 | 1.07 | 5.23 | 282.85 | 1.53 | 7.22 | 231.41 | 0.96 | 7.77 | *** | *** | *** |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Wyoming | WY | 225.38 | 0.93 | 3.52 | 274.78 | 0.91 | 1.79 | 223.20 | 1.38 | 4.14 | 276.69 | 1.18 | 3.96 |

**Table A3**
***NAEP Data for Grades 4 & 8—25th Percentile***

| | | | Grade 4 | | | | | | Grade 8 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1992 | | | 2000 | | | 1992 | | | 2000 | | |
| | State | Abbrev. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. |
| **Hi-stakes states** | **National (Public)** | **NAT** | **197.43** | **0.83** | **6.52** | **206.01** | **1.37** | **7.47** | **241.67** | **1.54** | **6.39** | **250.23** | **0.92** | **7.39** |
| | **Alabama** | **AL** | **186.19** | **1.56** | **4.51** | **198.99** | **1.57** | **5.78** | **227.94** | **1.94** | **5.45** | **238.66** | **1.61** | **4.94** |
| | **Georgia** | **GA** | **193.74** | **1.38** | **5.23** | **199.37** | **2.05** | **6.51** | **235.74** | **1.52** | **4.68** | **241.94** | **2.17** | **7.20** |
| | **Indiana** | **IN** | **201.76** | **1.35** | **3.32** | **217.40** | **1.84** | **6.78** | **247.81** | **1.45** | **4.58** | **262.95** | **1.46** | **7.30** |
| | **Louisiana** | **LA** | **183.26** | **2.61** | **4.00** | **199.27** | **1.44** | **7.68** | **227.41** | **2.61** | **4.27** | **236.46** | **2.19** | **5.76** |
| | **Maryland** | **MD** | **193.40** | **2.14** | **4.05** | **199.92** | **2.20** | **8.93** | **237.90** | **2.53** | **4.68** | **250.76** | **1.42** | **10.56** |
| | **Minnesota** | **MN** | **209.48** | **1.07** | **3.36** | **217.57** | **1.49** | **5.80** | **261.12** | **1.13** | **3.46** | **268.36** | **2.08** | **5.49** |
| | **Mississippi** | **MS** | **180.49** | **1.65** | **4.85** | **191.72** | **1.58** | **4.26** | **222.43** | **1.11** | **7.17** | **232.95** | **1.93** | **7.47** |
| | **New Mexico** | **NM** | **193.25** | **1.90** | **7.35** | **194.81** | **2.13** | **12.30** | **238.48** | **1.07** | **5.34** | **239.06** | **2.63** | **11.55** |
| | **New York** | **NY** | **197.68** | **0.90** | **5.29** | **207.19** | **2.74** | **11.53** | **241.93** | **2.65** | **8.49** | **254.36** | **2.50** | **13.13** |
| | **North Carolina** | **NC** | **190.45** | **1.32** | **3.86** | **214.50** | **1.29** | **13.34** | **234.58** | **1.53** | **3.26** | **256.92** | **2.03** | **13.85** |
| | **Ohio** | **OH** | **198.22** | **1.49** | **6.04** | **212.60** | **1.78** | **9.95** | **245.39** | **1.92** | **6.09** | **262.54** | **2.22** | **8.70** |
| | **South Carolina** | **SC** | **191.18** | **1.15** | **4.81** | **200.81** | **1.47** | **7.46** | **236.28** | **1.22** | **5.91** | **243.31** | **1.43** | **6.85** |
| | **Tennessee** | **TN** | **190.50** | **2.09** | **3.87** | **199.96** | **1.82** | **3.77** | **236.01** | **1.46** | **4.89** | **240.19** | **1.97** | **4.56** |
| | **Texas** | **TX** | **198.33** | **1.63** | **7.58** | **215.62** | **1.55** | **15.44** | **238.59** | **1.25** | **6.58** | **254.68** | **1.88** | **9.57** |
| | **Virginia** | **VA** | **199.36** | **1.56** | **5.24** | **212.13** | **1.41** | **10.79** | **244.00** | **1.75** | **5.28** | **254.04** | **1.43** | **9.97** |
| Lo-stakes states | National (Public) | NAT | 197.43 | 0.83 | 6.52 | 206.01 | 1.37 | 7.47 | 241.67 | 1.54 | 6.39 | 250.23 | 0.92 | 7.39 |
| | California | CA | 184.59 | 2.29 | 12.20 | 190.62 | 3.06 | 8.96 | 234.77 | 2.69 | 8.26 | 236.85 | 1.50 | 8.68 |
| | Kentucky | KY | 195.22 | 0.92 | 3.29 | 202.18 | 1.42 | 8.24 | 239.48 | 1.68 | 4.56 | 250.50 | 1.36 | 9.47 |
| | Massachusetts | MA | 207.20 | 1.52 | 6.99 | 217.83 | 1.52 | 10.44 | 249.66 | 2.04 | 8.14 | 262.67 | 1.68 | 12.13 |
| | Michigan | MI | 199.93 | 2.66 | 5.26 | 211.37 | 1.47 | 8.37 | 244.02 | 2.16 | 6.06 | 255.94 | 2.63 | 6.54 |
| | Missouri | MO | 203.07 | 1.44 | 4.37 | 211.80 | 1.17 | 9.71 | 249.52 | 1.83 | 4.47 | 253.71 | 2.14 | 8.58 |
| | Oklahoma | OK | 202.83 | 0.78 | 7.29 | 209.85 | 1.09 | 10.47 | 248.03 | 1.28 | 6.30 | 251.88 | 2.20 | 8.61 |
| | West Virginia | WV | 195.79 | 1.68 | 4.42 | 208.48 | 0.89 | 10.07 | 238.11 | 0.97 | 5.79 | 250.95 | 2.53 | 11.08 |
| | Arizona | AZ | 194.62 | 2.30 | 4.93 | 197.85 | 1.92 | 11.85 | 243.87 | 1.40 | 5.71 | 248.11 | 2.88 | 9.08 |
| | Arkansas | AR | 189.54 | 1.35 | 5.43 | 197.70 | 1.16 | 6.58 | 234.09 | 1.48 | 6.24 | 239.08 | 2.40 | 8.18 |
| | Connecticut | CT | 206.41 | 1.36 | 6.58 | 215.94 | 2.07 | 9.95 | 249.75 | 1.39 | 6.59 | 258.54 | 1.39 | 10.19 |
| | Hawaii | HI | 191.80 | 2.05 | 5.75 | 195.94 | 1.60 | 10.17 | 232.05 | 1.05 | 4.86 | 239.91 | 1.65 | 7.28 |
| | Idaho | ID | 203.53 | 1.90 | 3.45 | 210.44 | 1.47 | 5.91 | 255.49 | 0.94 | 3.17 | 257.78 | 1.29 | 4.76 |
| | Maine | ME | 213.90 | 1.71 | 5.66 | 213.44 | 1.85 | 10.12 | 259.10 | 1.14 | 4.43 | 263.10 | 1.80 | 8.56 |
| | Nebraska | NE | 205.72 | 1.56 | 4.13 | 206.20 | 2.61 | 7.58 | 257.28 | 1.35 | 4.07 | 260.51 | 1.92 | 3.50 |
| | North Dakota | ND | 211.92 | 1.05 | 1.74 | 214.13 | 1.49 | 5.72 | 264.64 | 1.25 | 2.40 | 264.75 | 1.85 | 3.80 |
| | Rhode Island | RI | 195.24 | 3.28 | 5.81 | 205.89 | 1.68 | 11.70 | 243.87 | 1.47 | 5.10 | 250.52 | 2.12 | 11.77 |
| | Utah | UT | 205.63 | 1.13 | 4.03 | 208.76 | 1.83 | 6.66 | 253.82 | 1.37 | 4.37 | 254.10 | 1.96 | 5.82 |
| | Wyoming | WY | 208.57 | 1.56 | 3.52 | 212.13 | 2.21 | 6.07 | 255.13 | 1.05 | 3.95 | 256.90 | 1.66 | 3.96 |

**Table A4**
***NAEP Data for Cohort Analysis—25th Percentile***

| | | | Cohort (1992 to 1996) | | | | | | Cohort (1996 to 2000) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Grade 4 (1992) | | | Grade 8 (1996) | | | Grade 4 (1996) | | | Grade 8 (2000) | | |
| | State | Abbrev. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. | Score | s.e. | % Excl. |
| **Hi-stakes states** | **National (Public)** | **NAT** | **197.43** | **0.83** | **6.52** | **246.71** | **1.32** | **4.84** | **201.49** | **1.31** | **6.36** | **250.23** | **0.92** | **7.39** |
| | **Alabama** | **AL** | **186.19** | **1.56** | **4.51** | **233.09** | **2.26** | **7.15** | **190.12** | **1.40** | **6.28** | **238.66** | **1.61** | **4.94** |
| | **Florida** | **FL** | **193.07** | **1.81** | **8.26** | **240.26** | **2.14** | **9.82** | **194.89** | **1.38** | **9.95** | **xxx** | **xxx** | **xxx** |
| | **Georgia** | **GA** | **193.74** | **1.38** | **5.23** | **237.74** | **3.33** | **6.99** | **195.38** | **1.97** | **7.23** | **241.94** | **2.17** | **7.20** |
| | **Indiana** | **IN** | **201.76** | **1.35** | **3.32** | **255.12** | **2.27** | **5.61** | **211.48** | **1.20** | **5.28** | **262.95** | **1.46** | **7.30** |
| | **Louisiana** | **LA** | **183.26** | **2.61** | **4.00** | **231.62** | **1.95** | **5.96** | **189.56** | **1.24** | **7.66** | **236.46** | **2.19** | **5.76** |
| | **Maryland** | **MD** | **193.40** | **2.14** | **4.05** | **242.34** | **2.28** | **6.67** | **197.44** | **1.74** | **7.68** | **250.76** | **1.42** | **10.56** |
| | **Minnesota** | **MN** | **209.48** | **1.07** | **3.36** | **262.11** | **1.48** | **2.96** | **214.60** | **2.36** | **6.12** | **268.36** | **2.08** | **5.49** |
| | **Mississippi** | **MS** | **180.49** | **1.65** | **4.85** | **227.84** | **1.54** | **6.72** | **188.23** | **1.17** | **5.97** | **232.95** | **1.93** | **7.47** |
| | **Nevada** | **NV** | **xxx** | **xxx** | **xxx** | **xxx** | **xxx** | **xxx** | **198.44** | **2.65** | **8.90** | **245.65** | **1.30** | **10.17** |
| | **New Mexico** | **NM** | **193.25** | **1.90** | **7.35** | **239.56** | **1.38** | **7.80** | **193.33** | **2.78** | **11.94** | **239.06** | **2.63** | **11.55** |
| | **New York** | **NY** | **197.68** | **0.90** | **5.29** | **247.24** | **2.73** | **7.63** | **203.19** | **2.47** | **7.80** | **254.36** | **2.50** | **13.13** |
| | **North Carolina** | **NC** | **190.45** | **1.32** | **3.86** | **243.79** | **1.43** | **4.45** | **203.95** | **1.96** | **6.89** | **256.92** | **2.03** | **13.85** |
| | **South Carolina** | **SC** | **191.18** | **1.15** | **4.81** | **237.73** | **1.78** | **5.79** | **192.69** | **1.88** | **5.65** | **243.31** | **1.43** | **6.85** |
| | **Tennessee** | **TN** | **190.50** | **2.09** | **3.87** | **240.50** | **1.66** | **4.33** | **199.29** | **1.87** | **6.40** | **240.19** | **1.97** | **4.56** |
| | **Texas** | **TX** | **198.33** | **1.63** | **7.58** | **246.64** | **1.92** | **8.64** | **208.78** | **2.16** | **10.32** | **254.68** | **1.88** | **9.57** |
| | **Virginia** | **VA** | **199.36** | **1.56** | **5.24** | **246.25** | **1.14** | **7.26** | **202.05** | **1.86** | **6.68** | **254.04** | **1.43** | **9.97** |
| Lo-stakes states | National (Public) | NAT | 197.43 | 0.83 | 6.52 | 246.71 | 1.32 | 4.84 | 201.49 | 1.31 | 6.36 | 250.23 | 0.92 | 7.39 |
| | Arizona | AZ | 194.62 | 2.30 | 4.93 | 246.13 | 1.41 | 8.52 | 197.92 | 1.20 | 12.42 | 248.11 | 2.88 | 9.08 |
| | Arkansas | AR | 189.54 | 1.35 | 5.43 | 238.89 | 1.81 | 7.01 | 195.26 | 1.65 | 6.69 | 239.08 | 2.40 | 8.18 |
| | California | CA | 184.59 | 2.29 | 12.20 | 237.10 | 2.35 | 9.99 | 186.48 | 1.57 | 15.75 | 236.85 | 1.50 | 8.68 |
| | Colorado | CO | 201.45 | 1.01 | 5.24 | 253.51 | 1.30 | 4.41 | 206.62 | 1.41 | 8.38 | xxx | xxx | xxx |
| | Connecticut | CT | 206.41 | 1.36 | 6.58 | 256.53 | 1.23 | 8.28 | 213.74 | 1.67 | 8.09 | 258.54 | 1.39 | 10.19 |
| | Delaware | DE | 195.68 | 1.74 | 5.30 | 243.40 | 1.81 | 8.64 | 193.38 | 1.89 | 7.04 | xxx | xxx | xxx |
| | Hawaii | HI | 191.80 | 2.05 | 5.75 | 237.88 | 1.01 | 5.21 | 192.54 | 2.13 | 5.78 | 239.91 | 1.65 | 7.28 |
| | Iowa | IA | 211.76 | 1.29 | 3.26 | 264.42 | 0.98 | 5.23 | 212.98 | 1.37 | 5.62 | xxx | xxx | xxx |
| | Kentucky | KY | 195.22 | 0.92 | 3.29 | 245.60 | 1.36 | 4.69 | 200.64 | 2.61 | 5.72 | 250.50 | 1.36 | 9.47 |
| | Maine | ME | 213.90 | 1.71 | 5.66 | 264.62 | 1.07 | 4.71 | 214.34 | 1.16 | 7.52 | 263.10 | 1.80 | 8.56 |
| | Massachusetts | MA | 207.20 | 1.52 | 6.99 | 254.60 | 2.71 | 8.02 | 210.79 | 2.03 | 9.02 | 262.67 | 1.68 | 12.13 |
| | Michigan | MI | 199.93 | 2.66 | 5.26 | 253.20 | 2.35 | 5.12 | 206.67 | 1.34 | 6.17 | 255.94 | 2.63 | 6.54 |
| | Missouri | MO | 203.07 | 1.44 | 4.37 | 251.99 | 1.81 | 7.10 | 206.18 | 1.72 | 4.94 | 253.71 | 2.14 | 8.58 |
| | Montana | MT | xxx | xxx | xxx | 262.24 | 1.52 | 3.22 | 209.86 | 2.15 | 4.77 | 267.22 | 1.99 | 5.44 |
| | Nebraska | NE | 205.72 | 1.56 | 4.13 | 262.92 | 1.49 | 4.33 | 208.87 | 1.85 | 5.02 | 260.51 | 1.92 | 3.50 |
| | North Dakota | ND | 211.92 | 1.05 | 1.74 | 264.60 | 1.28 | 3.32 | 214.20 | 1.39 | 3.56 | 264.75 | 1.85 | 3.80 |
| | Oregon | OR | xxx | xxx | xxx | 253.60 | 2.33 | 3.92 | 203.69 | 1.41 | 8.77 | 258.06 | 1.13 | 6.19 |
| | Rhode Island | RI | 195.24 | 3.28 | 5.81 | 245.93 | 1.77 | 7.18 | 201.53 | 1.39 | 6.16 | 250.52 | 2.12 | 11.77 |
| | Utah | UT | 205.63 | 1.13 | 4.03 | 256.84 | 1.48 | 6.06 | 208.14 | 2.05 | 5.79 | 254.10 | 1.96 | 5.82 |
| | Vermont | VT | xxx | xxx | xxx | 258.86 | 1.45 | 4.29 | 205.96 | 1.67 | 6.08 | 262.27 | 1.44 | 9.54 |
| | West Virginia | WV | 195.79 | 1.68 | 4.42 | 243.68 | 1.22 | 8.49 | 204.10 | 0.79 | 8.34 | 250.95 | 2.53 | 11.08 |
| | Wisconsin | WI | 210.39 | 0.47 | 5.23 | 262.07 | 2.25 | 7.22 | 213.19 | 0.90 | 7.77 | xxx | xxx | xxx |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Wyoming | WY | 208.57 | 1.56 | 3.52 | 255.73 | 1.22 | 1.79 | 204.54 | 1.74 | 4.14 | 256.90 | 1.66 | 3.96 |