High-Stakes Testing and Student Achievement: Problems for the No Child Left Behind Act

by

Sharon L. Nichols
Assistant Professor
University of Texas at San Antonio

Gene V Glass
Regents' Professor
Arizona State University

David C. Berliner
Regents' Professor
Arizona State University

The Great Lakes Center for Education Research & Practice
PO Box 1263
East Lansing, MI 48826
Phone: (517) 203-2940
Email: greatlakescenter@greatlakescenter.org
Web Site: http://www.greatlakescenter.org

This research was made possible by a grant from the Great Lakes Center for Education Research and Practice.

TABLE OF CONTENTS

High-Stakes Testing and Student Achievement:

Problems for the No Child Left Behind Act

Sharon L. Nichols
University of Texas at San Antonio

Gene V Glass
Arizona State University

David C. Berliner
Arizona State University

Executive Summary

Under the federal No Child Left Behind Act of 2001 (NCLB), standardized test scores are the indicator used to hold schools and school districts accountable for student achievement. Each state is responsible for constructing an accountability system, attaching consequences—or stakes—for student performance. The theory of action implied by this accountability program is that the pressure of high-stakes testing will increase student achievement. But this study finds that pressure created by high-stakes testing has had almost no important influence on student academic performance.

To measure the impact of high-stakes testing pressure on achievement and to account for the differences in testing pressure among the states, researchers created the Pressure Rating Index (PRI). The PRI was used in two ways. Correlations between the PRI and National Assessment for Educational Progress (NAEP) results from 1990 to 2003 in 25 states were analyzed and the PRI was used in replications of previous research. These analyses revealed that:

- States with greater proportions of minority students implement accountability systems that exert greater pressure. This suggests that any problems associated with high-stakes testing will disproportionately affect America's minority students.

- High-stakes testing pressure is negatively associated with the likelihood that eighth and tenth graders will move into 12th grade. Study results suggest that increases in testing pressure are related to larger numbers of students being held back or dropping out of school.

- Increased testing pressure produced no gains in NAEP reading scores at the fourth- or eighth-grade levels.

- Prior increases in testing pressure were weakly linked to subsequent increases in NAEP math achievement at the fourth-grade level. This finding emerged for all ethnic subgroups, and it did not exist prior to 1996. While the authors believe a causal link exists between earlier pressure increases and later fourth-grade math achievement increases, they also point out that math in the primary grades is far more standardized across the country than the math curriculum in middle school and, therefore, drilling students and teaching to the test could have played a role in this increase**.** This interpretation is supported by the lack of evidence that earlier pressure increases produced later achievement increases for eighth-grade math achievement or for fourth- and eighth-grade reading achievement.

The authors conclude that there is no convincing evidence that the pressure associated with high-stakes testing leads to any important benefits for students' achievement. They call for a moratorium on policies that force the public education system to rely on high-stakes testing.

Introduction

Supporters of the practice of high-stakes testing believe that the quality of American education can be vastly improved by introducing a system of rewards and sanctions for students' academic performance.[1] When faced with large incentives and threatening punishments, administrators, teachers, and students, it is believed, will take schooling more seriously and work harder to obtain rewards and avoid humiliating punishments. But educators and researchers have argued that serious problems accompany the introduction of high-stakes testing. Measurement specialists oppose high-stakes testing because using a single indicator of competence to make important decisions about individuals or schools violates the professional standards of the measurement community.[2] Other critics worry that the unintended effects of high-stakes testing not only threaten the validity of test scores, but also lead to "perverse"[3] and "corrupt" educational practice.[4] Teachers report that the pressure of doing well on a test seriously compromises instructional practice.[5] And still others worry that the exaggerated pressure on students and teachers to focus on test preparation is thwarting teachers' intentions to care for students' needs apart from those that lead to the scores they receive on examinations.[6] It is also argued by many that the measurement systems we currently have cannot support the demands of those who make educational policy.[7]

The assumption embedded in the current promotion of a high-stakes accountability model of education is that students and teachers need to work harder and that by pressuring them with the threat of sanctions and enticing them with financial incentives, they would expend more effort and time on academic pursuits, and thus

learning would increase. This rationale is problematic for several reasons. Learning is a complicated endeavor and as most educators would argue, extrinsic rewards alone cannot overcome the range of background experiences and individual differences in learning and motivation students bring to school.[8] Still, with significant bipartisan support, the passage of the No Child Left Behind Act (NCLB) of 2001 instantiated this notion of academic accountability in education—at least for now. But, is it working? Does the threat of rewards and sanctions increase achievement?

Although the literature on the mostly harmful and *un*intended effects of high-stakes testing is growing rapidly,[9] existing research on the relationship between high-stakes testing and its *intended* impact on achievement is mixed and inconclusive. Some studies find no evidence that high-stakes testing impacts achievement.[10] Others argue that the data for or against are not sufficiently robust to reject outright the use of high-stakes testing for increasing achievement.[11] And others report mixed effects, finding high-stakes testing to be beneficial for certain student groups but not others.[12]

One potential explanation for the mixed conclusions about the effectiveness of high-stakes testing on achievement could lie in measurement differences in the characterization of a high-stakes testing state (i.e., which states truly have high-stakes and which only appear to have them?). Some researchers study the issue using a two-group comparison—analyzing achievement trends in states with high-stakes testing policies against those without.[13] Others have studied the issue by rating states along a continuum of low- to high-stakes state (i.e., a low-stakes state has fewer consequences for low performance than a high-stakes state). As more states implement high-stakes testing, the

rating measurement approach becomes more important than the two-group comparison approach. Exploring new measurement methods is one goal of this report.

This study adds to the literature in two important ways. First, we employ qualitative and quantitative methods to measure the pressure on teachers, students, and parents exerted by a "high-stakes testing" system. An examination of the research on accountability implementation both before and after NCLB was signed into law uncovered the inadequacies of existing measurement approaches for capturing the range of pressures that high-stakes testing exerted on students and educators or at the very least, they showed little agreement from one study to the next. Thus, a significant goal of this study is to create a more valid system for measuring the pressure that high-stakes testing systems apply to educators and their students. Our second goal is to use this newly created rating system to conduct a series of analyses to examine whether the practice of high-stakes testing increases achievement. This is addressed in two ways. First, findings from research by Carnoy and Loeb[14] (whose recent report concluded that strength of a state's accountability model is related to math achievement gains, specifically for minority students and for eighth graders), are challenged. This research replicates their analyses, but replaces their high-stakes pressure index with ours. Second, a series of analyses to investigate the relationship between high-stakes testing implementation and achievement trends over time are computed.

Why High-Stakes Testing?

The publication of *A Nation at Risk*[15] alarmed citizens with its claim that the American public education system was failing. As the report noted, it was believed that

if the education system did not receive a major overhaul, our economic security would be severely compromised. American culture has internalized this claim to such a degree that questions about how to solve this "crisis" continue to be at the top of many policy makers' agendas. Although our education system is not as bad off as some would have the public believe,[16] the rhetoric of a failing education system has led to a series of initiatives that have transformed the role and function of the American public school system. High-stakes testing holds a prominent place in this transformation.

The earliest and most common form of high-stakes testing was the practice of attaching consequences to high school graduation exams (i.e., students had to pass a test to receive a high school diploma). New York's Regents examinations served this purpose for over 100 years[17] and states such as Florida, Alabama, Nevada, and Virginia had instituted high-stakes graduation exams at least as far back as the early to mid 1980s.[18] But in the years since *A Nation at Risk,* the rhetoric of high expectations, accountability, and ensuring that all students—especially those from disadvantaged backgrounds—have an equal opportunity to receive quality education has been accompanied by a series of federal initiatives including Clinton's 1994 re-authorization of the 1965 Elementary and Secondary School Act, subsequent education "policy summits," and George H. W. Bush's Goals 2000. In combination, these initiatives have progressively increased the demands on teachers and their students and have laid the groundwork for what was to come next—an unprecedented federal intervention on state-level education policy making[19] that directs all states toward a single goal (i.e., 100 percent of students reaching "proficiency") via a single system of implementation (i.e., standards-based assessment and accountability).

No Child Left Behind: Changing the Landscape of Accountability

The construction and passage of the No Child Left Behind Act (NCLB) occurred under the leadership of Rod Paige and George W. Bush. In Texas, in the decade before they went to Washington, Bush as governor and Paige as superintendent of Houston school district had built and implemented a controversial high-stakes accountability system that placed increasing demands and expectations on students for well over a decade. And while other states were also implementing accountability systems (Kentucky and New York among others), Texas's "success" of holding students and educators accountable for learning was quite visible. Although the "myth" of Texas's success has been critically examined and documented,[20] it was too late (or more likely, no one paid close attention) and NCLB, influenced by the programs implemented in Texas and elsewhere was passed in 2001 and signed into law on January 8, 2002.[21]

The goal of NCLB was ambitious—to bring all students up to a level of academic "proficiency" within a 15-year period. As of the day it was signed into law, states had to initiate a strategic plan for meeting the range of assessment and accountability provisions the law mandated. States that did not were threatened by the loss of billions in Title I funding (see Table 1 for an overview of the law's major mandates). At the core of these mandates is that states adopt a system of accountability defined by sanctions and rewards that would be applied to schools, teachers, and students in the event they did not meet pre-defined achievement goals (see Table 2 for an outline of NCLB-defined rewards and sanctions).

Table 1: Overview of Requirements for States Under NCLB

1.	All states must identify a set of academic standards for core subject areas at each grade level;
2.	States must create a state assessment system to monitor student progress toward meeting these state-defined standards;
3.	States must require schools and districts to publish report cards identifying academic achievement of its students in aggregate and disaggregated by ethnicity and other sub groups (e.g., for racial minorities, students for whom English is a Second Language (ESL) and special education students);
4.	States must create a system of labels that communicate to the community how local schools and districts are performing;
5.	States must create a plan (i.e., Adequate Yearly Progress or AYP) that would ensure 100 percent of its students will reach academic proficiency by the year 2014-2015; and
6.	States must come up with a system of accountability that includes rewards and sanctions to schools, educators, and students that are tied to whether they meet state's goals outlined in the AYP plan.

Source: No Child Left Behind Act (NCLB) of 2001 § 1001, 20 U.S.C. § 6301. Retrieved February 18, 2005, from: http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf

The law is massive and forces states to allocate significant resources in the form of time, energy, and especially money towards its implementation—implementation that has been especially cumbersome[22] if not potentially counterproductive to the goals of schooling.[23] Most states were not ready to comply with the range of demands from NCLB. Some didn't have any sort of assessment system in place, whereas others were just beginning to pilot theirs. Similarly, some states were already holding students and their teachers accountable, whereas others had no plans or intentions of doing so. The demands associated with NCLB have caused problems and challenges for many states. In the first two to three years of implementation, most states have experienced significant financial and logistical barriers in implementing two of the primary accountability provisions stipulated under NCLB: provision of supplementary services and allowing

students to transfer out of "under performing" schools.[24] And, in many cases, the demands of the law have been met with negativity by those it arguably impacts the most—teachers.[25] Ultimately, the pace at which states are able to enact and implement the range of accountability provisions outlined by NCLB varies a great deal. It is this incredible range of accountability implementation that makes the study of high-stakes testing impact more complicated, but it is this complexity that is addressed by this study.

Table 2: NCLB Sanction and Reward Guidelines

Sanctions
1. Schools failing to meet adequate yearly progress (AYP) for two consecutive years must be identified as needing improvement. Technical assistance is to be provided and public school choice offered;
2. Schools failing to meet AYP for three years must offer pupils from low-income families the opportunity to receive instruction from supplemental services, (plus corrective actions in #1 above);
3. Schools failing to meet AYP for four consecutive years must take one of the following specified "corrective actions."
a. Replacing school staff, appointing outside expert to advise school, extend school day or year, change school internal organization structure (plus corrective actions in 1 and 2 above).
4. Schools that fail to meet AYP for five consecutive years must be "restructured." Such restructures must consist of one or more of the following actions:
a. reopening as a charter school, replacing all or most school staff, state takeover or school operations (if permitted under state law), or other major restructuring of school governance (plus 1-3 above).
Rewards
1. States must develop strategies related to high performing schools, or those showing improvement such as:
a. Academic achievement Awards: Receiving recognition when they close the achievement gap; or when they exceed AYP for two consecutive years.
b. "Distinguished schools" designations: identifying those schools that have made the greatest gains as "models" for low-performing schools.
2. Financial awards to teachers in schools that have made the greatest gains.

Source: No Child Left Behind Act (NCLB) of 2001 § 1001, 20 U.S.C. § 6301. Available online, accessed February 18, 2005, from, http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf

High-Stakes Testing and Achievement

A series of studies have emerged attempting to examine the effects of high-stakes testing on student achievement. Amrein and Berliner, Rosenshine, and Braun[26] debated the merits of high-stakes testing for improving achievement, often locating their conflicting conclusions in the statistical analyses they applied. Amrein and Berliner used time trend analysis to study the effectiveness of high-stakes testing on achievement at both the K-8 and high school levels. They analyzed achievement trends across time in high-stakes testing states against a national average. Their extensive and descriptive set of results are organized by state for which they noted whether there was "strong" or "weak" evidence to suggest whether achievement had "increased" or "decreased" in fourth- and eighth-grade National Assessment of Educational Progress (NAEP) scores in math and reading. They concluded that "no consistent effects across states were noted. Scores seemed to go up or down in random pattern after high-stakes test were introduced, indicating no consistent state effects as a function of high-stakes testing policy."[27]

In a reanalysis of the data addressing what were viewed as flaws in Amrein and Berliner's method and design—namely a lack of control group—Rosenshine found that average NAEP increases were greater in states with high-stakes testing polices than those in a control group of states without. Still, when he disaggregated the results by state, Rosenshine concluded that "although attaching accountability to statewide tests worked well in some high-stakes states it was not an effective policy in all states."[28] Again, no consistent effect was found.

In a follow-up response to Rosenshine, Amrein-Beardsley and Berliner [29] adopted his research method using a control group to examine NAEP trends over time, but they

also included in their analysis NAEP exclusion rates.[30] They concluded that although states with high-stakes tests seemed to outperform those without high-stakes tests on the fourth grade math NAEP exams, when controlling for exclusion rates, they found that this difference disappeared. They argued NAEP achievement in high-stakes testing states is likely to be inflated by the exclusion of greater numbers of lower achieving students.

Braun also critiqued Amrein and Berliner on methodological grounds. In his analysis of fourth- and eighth-grade math achievement (he did not look at reading) across the early 1990s, he found that when standard error estimates are included in the analyses, NAEP gains were greater in states with high-stakes testing than in those without, in spite of exclusion rate differences. He concludes, "The strength of the association between states' gains and a measure of the general accountability efforts in the states is greater in the eighth grade than in the fourth."[31] However, in a separate analysis following cohorts of students (1992 fourth-grade math and 1996 eighth-grade math; 1996 fourth-grade math and 2000 eighth-grade math), he found that high-stakes testing effects largely disappeared. As students progress through school, there is no difference in achievement trends between states with high-stakes testing and those without. His conclusions about usefulness of high-stakes testing as a widespread policy are tentative. "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."[32]

Carnoy and Loeb provide yet another set of analyses to describe the impact of high-stakes testing using a completely different approach for measuring accountability and focusing on effects by student ethnicity. In contrast to others who adopted Amrein

and Berliner's initial categorization, Carnoy and Loeb operationalize "high-stakes testing" in terms of the "strength" of the accountability in each state, rating each state on a 5-point scale to perform a series of regression analyses. Their analysis leads them to conclude that accountability strength is significantly related to math achievement gains among eighth graders, especially for African American and Hispanic students.

Carnoy and Loeb also consider the relationship between students' grade-to-grade progression rates with strength of accountability. Others have argued that high-stakes testing influences a greater number of students, especially minority students, to drop out of school.[33] Carnoy and Loeb found no relationships between accountability strength and student progression rates.

Conclusions From the Research

To date there is no consistent evidence that high-stakes testing works to increase achievement. Although data suggest the possibility that high-stakes testing affects math achievement—especially among eighth graders and for some sub-groups of students—the findings simply are not sufficiently consistent to make the stronger claim that math learning is benefited by high-stakes testing pressure. Part of the concern is that it cannot be determined definitively whether achievement gains on state assessments are real or whether they are the outcome of increased practice and teaching to the test. That is why National Assessment of Educational Progress or other measures of student learning are needed. Thus, in spite of the claims of some who seem to argue that the benefits of high-stakes testing are well established,[34] it appears that more empirical studies are needed to

determine whether high-stakes testing has the intended effect of increasing student learning.

Measuring High-Stakes Testing Pressure

In this section, we describe our approach to measuring high-stakes testing pressure. Previous researchers studying the relationship between high-stakes testing pressure and achievement have differed significantly in their categorization of states with respect to the pressure on teachers and students exerted by the accountability programs. This section begins with a brief overview of existing systems followed by a detailed overview of the methods adopted to measure pressure across our study states.

Existing Systems

Amrein and Berliner studied high-stakes testing impact by identifying the timing and nature of each state's high-stakes policies and comparing their achievement trends against a national average. Others following Amrein and Berliner's categorization of high- versus low-stakes states conducted "cleaner" two group comparisons to study achievement patterns in high-stakes testing states against those without high-stakes testing systems.[35] But, the rapidly increasing number of states joining the list of those with high-stakes testing—and the implementation of No Child Left Behind (NCLB)—has made a two-group design far less useful.

Other approaches characterize accountability implementation and impact with a numerical index, rating states along a continuum that is defined by some aspect of accountability. Swanson and Stevenson[36] crafted an index of "policy activism" that measured the degree to which states were implementing any one of 22 possible state

policy activities related to standards-based assessment and accountability. These 22 activities were organized into four categories: (a) content standards, (b) performance standards, (c) aligned assessments, and (d) professional standards. States received one of three scores across all 22 possible policy activities (0=does not have a policy, 1=developing one, and 2=has enacted such a policy as of 1996) yielding a state-level index of overall "policy activism" (scale ranged from -1.61 to 2.46). Swanson and Stevenson's index measures the relative amount of standards-based reform activity as of 2001.

Carnoy and Loeb created an index-like system, but one that measured each state's accountability "strength." They noted, "The 0-5 scale captures degrees of state external pressure on schools to improve student achievement according to state-defined performance criteria."[37] Thus, their index was crafted to represent a hypothetical degree of "pressure" on teachers and students to perform well on state tests. This pressure is based on (a) the grades in which students were tested, (b) school accountability, (c) repercussions for schools, (d) strength of repercussions for schools, (e) if there is a high school exit test (in 2000), and if so, the grade at which first high school test is given, and (f) the first class that had to pass the test to get a diploma (all information based on data as of 1999-2000).[38] Although they provide a general description of what each index value represents, their overall rationale is vague. For example, to receive the highest strength of accountability score they note, "States receiving a 5 had to have students tested in several different grades, schools sanctioned or rewarded based on student test scores, and a high school minimum competency test required for graduation. Other states

had some of these elements, but not others."[39] Carnoy and Loeb provide very little information as to how they differentiated a 5 score versus a 4 score and so on.

Lastly, researchers from Boston College came up with a three by three matrix of accountability where one dimension is defined by the severity of the consequences to students (high, moderate, low) and the other by the severity of consequences to teachers, schools, and districts (again, high, moderate, or low).[40] Each state receives one of nine possible characterizations to describe overall amount of pressure as it relates to adults versus students (H/H, L/L, etc.). Nominal representations were transposed into an ordinal-like rating to calculate possible overlap among the existing systems for measuring high-stakes pressure.

The ratings assigned by the various systems of Amrein and Berliner, Swanson and Stevenson, Pedulla et al., and Carnoy and Loeb are displayed in Table 3 followed by a table of correlations (Table 4). Note that Amrein and Berliner's rating was based on the number of stakes identified in their initial report.[41] Carnoy and Loeb (in a cautious acknowledgement of the ambiguities in any rating scale) assigned two different ratings for four states (California, Maryland, New Mexico, and New York). Both rating scales are included here. The Boston College classification was converted into two numerical classification systems. The Education Commission of the States (ECS)[42] rating system was based on a tally of the number of potential sanctions on the law books as of 2001.[43]

Table 3: Outline of Existing Rating Systems

	Amrein & Berliner	Policy Activism	Carnoy 1	Carnoy 2	Boston Rating 1*	Boston Rating 2**	ECS
Alabama	4	2.195	4	4	4	9	4
Alaska	0	-0.949	1	1	4	6	0
Arizona	0	-0.395	2	2	2	6	2
Arkansas	0	-0.270	1	1	1	8	1
California	5	0.090	4	2	4	9	2
Colorado	5	0.662	1	1	3	7	7
Connecticut	0	1.290	1	1	1	8	1
Delaware	6	0.206	1	1	5	9	2
Florida	5	-0.268	5	5	5	9	3
Georgia	1	0.660	2	2	3	9	3
Hawaii	0	0.320	1	1	1	4	1
Idaho	0	-0.268	1	1	3	3	0
Illinois	0	0.320	2.5	2.5	4	8	5
Indiana	4	0.899	3	3	5	9	2
Iowa	0	-1.606	0	0	1	1	3
Kansas	0	0.320	1	1	3	7	5
Kentucky	4	1.970	4	4	4	7	4
Louisiana	5	-0.030	3	3	3	9	3
Maine	0	1.290	1	1	1	7	1
Maryland	5	2.460	4	5	4	9	5
Massachusetts	3	0.320	2	2	4	9	2
Michigan	5	0.434	1	1	4	8	3
Minnesota	1	-0.395	2	2	4	6	1
Mississippi	2	0.550	3	3	3	9	3
Missouri	1	1.020	1.5	1.5	1	7	1
Montana	0	-1.261	1	1	2	4	0
Nebraska	0	-1.606	0	0	2	7	0
Nevada	4	0.320	1.5	1.5	5	9	2
New Hampshire	0	1.153	1	1	2	4	0
New Jersey	3	-0.395	5	5	5	9	2
New Mexico	5	0.780	4	5	4	9	4
New York	4	0.090	5	2	5	9	2
North Carolina	6	1.600	5	5	5	9	5
North Dakota	0	-0.026	1	1	2	4	0
Ohio	5	1.153	3	3	4	6	2
Oklahoma	2	0.434	1	1	3	7	8
Oregon	0	0.662	2.5	2.5	3	5	1
Pennsylvania	3	-0.661	1	1	4	8	2
Rhode Island	0	0.090	1	1	4	7	1

South Carolina	5	0.900	3	3	3	7	3
South Dakota	0	-0.802	1	1	2	4	0
Tennessee	4	0.320	1.5	1.5	3	9	3
Texas	6	-0.660	5	5	5	9	5
Utah	0	1.150	1	1	1	4	1
Vermont	0	-0.268	1	1	3	7	5
Virginia	2	0.550	2	2	1	9	1
Washington	0	0.206	1	1	4	6	0
West Virginia	3	0.900	3.5	3.5	3.5	8	3.5
Wisconsin	0	-0.395	2	2	4	6	0
Wyoming	0	-0.950	1	1	1	4	1

* where H/H = 5; H/M or M/H =4; H/L or L/H=3; M/L or L/M=2; and L/L=1
** where H/H=9; H/M=8; H/L=7; M/H=6; M/M=5; M/L=4; L/H=3; L/M=2; L/L=1

Table 4: Correlations of Existing Accountability Measurement Systems

	Amrein & Berliner	**Policy Activism**	**Carnoy 1**	**Carnoy 2**	**Boston Rating 1***	**Boston Rating 2****	**ECS**
Amrein & Berliner	1.000						
Policy Activism	0.361	1.000					
Carnoy 1	0.663	0.370	1.000				
Carnoy 2	0.636	0.433	0.926	1.000			
Boston Rating 1*	0.646	0.118	0.616	0.564	1.000		
Boston Rating 2**	0.655	0.361	0.575	0.541	0.561	1.000	
ECS	0.513	0.338	0.358	0.407	0.329	0.422	1.000

* where H/H = 5; H/M or M/H =4; H/L or L/H=3; M/L or L/M=2; and L/L=1
** where H/H=9; H/M=8; H/L=7; M/H=6; M/M=5; M/L=4; L/H=3; L/M=2; L/L=1

Amrein and Berliner, Carnoy and Loeb, and the Boston systems were all positively correlated in spite of being based on relatively different conceptualizations of accountability "strength." Nonetheless, the differences among these systems are great enough as to raise concern and focus attention on better ways of measuring high-stakes pressure. The policy activism scale is also positively related with other systems, suggesting some overlap between strength of accountability and degree to which policies are created and acted upon.

The Present Definition of High-Stakes Testing

As was the case with Carnoy and Loeb, the feature of high-stakes testing that we wish to capture in our measure is the construct of "pressure" as it relates to the amount of "press" or "threat" associated with performance on a particular test. However, our measurement approach to capturing this "threat" or "pressure" is based on a more differentiated conceptualization of high-stakes testing policy, practice, and implementation than has heretofore been carried out. Although laws and regulations provide a political description of accountability in each state, they cannot fully describe the level, nature, and extremely varied impact of the laws on individuals. For example, it might be state law to hold students back if they fail end-of-year exams, but the actual "threat" of this consequence as it is experienced by students, teachers, and parents depends on a great many influences including historical precedence (have students already been held back thus making the probability of it happening more realistic?), and the weight assigned to test performance (does a single test determine retention or are other considerations taken into account?). This type of differentiation is significant in terms of the actual pressure experienced by students and teachers.

In our measure of high-stakes pressure, state-level variation in high-stakes testing is accounted for by including both the actual laws as well as a proxy for their relative impact and implementation. The range of potential sanctions and rewards that could exist at the state level was identified first. For example, "Is it legal/mandatory for states to take over chronically underperforming schools?" and/or "Can states fire a teacher who works in a chronically underperforming school?" and so forth, using lists created by others as a starting point.[44] Once possible accountability laws were identified (see Table

5 for an overview), further aspects of the impact of the law were explored with follow-up questions such as: "Has the state ever taken over a school?" "How close are schools to be taken over?" "How much support does the state provide to schools to avoid this consequence?" "Do teachers accept the legitimacy of this potential sanction?" "If schools have been taken over, who assumed leadership?" "How successful was the transition?" To answer these questions, we (a) interviewed state department of education representatives, (b) consulted media sources, and (c) corresponded with ECS representatives who had access to a wide range of legal information.

Table 5: Overview of Possible Stakes

	Possible Sanctions	Possible Rewards
Consequences to:		
Districts	**Publicly Labeled "Failing"**	**Public Recognition**
	State Intervention	**Financial Rewards**
	State Takeover/Reorganization	
School Board	**Removal From Office**	**Public Recognition**
	Salary Reduced/Eliminated	
Administrators: Principals and Superintendents	**Publicly Associated With "Failing" School/District**	**Publicly Praised for Success**
	Salary Reduction	**Financial Bonuses**
	Termination From Job	
Schools	**Publicly Labeled as a Failing School, or One That "Needs Improvement"**	**Publicly Praised for Success**
	Financial Burden:	**Financial Rewards/Bonuses**
	• **paying to send students to go to another school**	
	• **state makes firing decisions**	
	• **pay to set up tutoring**	
	State Intervenes:	
	• **sends "improvement team" to evaluate school**	
	• **state makes firing decisions**	
	• **state turns school over to independent agency**	
	• **state takes over school**	
	• **state closes school**	
Teachers	**Publicly Labeled:**	**Publicly Praised for Success**
	• **bad teacher**	**Receive Financial Bonuses**
	• **associated with "failing" school**	
	Stricter Monitoring of Teaching	
	Job Loss	
Students	**K-8: Grade Retention**	**K-8: Parties Celebrating Test**
	High School: Diploma Withheld	**High School: College Scholarships**

Measurement Part I: Creating a Pressure Rating Index

The process of creating an index that would rank all 25 states[45] in this study based on a continuum of "pressure" associated with the practice of high-stakes testing is described in two main sections below. Part I includes a description of (a) the construction of portfolios used to tell the story of state-level accountability, (b) the procedures used to convert the portfolios into a pressure rating index (PRI), and (c) the validity analysis associated with this index. In part II, the procedures used to apply this index of state accountability pressure across time (1985-2004) are described.

Portfolios

The determination of a "pressure rating index" relied on a set of portfolios constructed to describe in as much detail as possible the past and current assessment and accountability practices of each state. These portfolios were crafted to tell the "story" of accountability; therefore, they include a wide range of documentation describing the politics and impact of a state's high-stakes testing program. All portfolios included three main sections: (a) an introduction essay, (b) a rewards/sanction sheet, (c) and newspaper stories. These are described in more detail next.

Context for Assessing State-Level Stakes

The first document in each portfolio was a summary essay of the state's past and current assessment and accountability plan (see Appendix A for an example). These essays included (a) some background information (e.g., name of past and current assessment system, implementation strategies), (b) a description of the most current assessment system, and (c) a summary of the rewards and sanctions (e.g., the current and past laws). The summary was written to be accessible to readers with a reasonable

acquaintance with schools and education more broadly and provided a "soft" introduction to the nature of assessment and accountability in each state. Importantly, these descriptions were informal and were not intended to represent fully the current or historical assessment and accountability activities in the state. Rather the goal of this initial portfolio document was to contextualize that state's accountability plan.

Rewards/Sanction Worksheet

Each portfolio contained a table that presented a range of questions and answers about what the state can do legally by way of consequences to districts, schools, and students (see Table 6 for an overview of all questions). This table drew heavily on data compiled by the Education Commission of States as of 2002 that described many of the accountability laws on state books as of 2001.[46]

In an effort to accurately represent a state's high-stakes testing environment, the rewards/sanctions worksheet was included to provide more detailed information about not only what is legally possible, but in what ways the law is viewed or implemented. For example, it might be the case that a teacher can be fired legally, but in reality a state may never have done this. This contrasts with another state where firing a teacher might not only be legal, but the state has already enacted the law and fired some teachers (An example of a completed rewards/sanctions worksheet is provided in Appendix B).

Table 6: Summary of Sanctions/Rewards Worksheet Questions

SANCTIONS	
Districts	
	Does the state have authority to put school districts on probation?
	Can the state remove a district's accreditation?
	Can the state withhold funding from the district?
	Can the state reorganize the district?
	Can the state take over the district?
	Does the state have the authority to replace superintendents?
Schools	
	Can schools be placed on probation?
	Can the state remove a school's accreditation?
	Can the state withhold funding from the school?
	Can the state reconstitute a school?
	Can the state close a school?
	Can the state take over a school?
	Does the state have the authority to replace teachers?
	Does the state have the authority to replace principals?
Students	
	K-8: Is grade to grade promotion contingent on exam?
	K-8: If yes, for students in what grades? And what is the timing of implementation?
	HIGH SCHOOL: Do students have to pass an exam in order to receive a diploma?
	HIGH SCHOOL: Are there alternative routes to receiving a diploma?
	HIGH SCHOOL: Are students required to attend remediation programs if they fail? (Who pays for it)?
	Students for whom English is a Second Language (LEP)
	Students with Disabilities
REWARDS	
Districts	
	Are districts rewarded for student performance?
	What types of awards are given (public recognition, certificates, monetary)?
	On what are rewards based (absolute performance or improvement)?
Schools	
	Are schools rewarded for student performance?
	What types of awards are given (public recognition, certificates, monetary)?
	On what are rewards based (absolute performance or improvement)?
Students	
	Are monetary awards or scholarships for college tuition given to high performing students?
	Public recognition of high performing students?

NOTE: Italicized statements are questions/considerations that were added for this project and were not part of the original ECS report.

Media

A range of newspaper stories were selected for inclusion in each portfolio. These newspaper articles added to the state's accountability "story" by providing a fuller picture of high stakes-testing impact. For example, a chronology of media coverage can add texture and context to what is known about state laws by describing (a) the timing of accountability implementation (e.g., when students were being tested, how well they did, who was doing well, who was doing poorly), (b) the general reaction to the accountability implementation (e.g., Were there many debates? Large agreement? Was it phased in slowly? Quickly?), and (c) editorials (e.g., op-ed pieces and letters to the editor) documenting readers' and/or experts' reactions to the accountability program. A full description of the strategy for selecting newspaper stories is described in Appendix F.

Media documentation was included because it provides a description of local cultural norms. Its value has been noted by others. "Documents are studied to understand culture—or the process and the array of objects, symbols, and meanings that make up social reality shared by members of a society."[47] Newspapers hold special relevance for representing cultural perspectives. Although they are not error or bias free, they contribute substantially to our shared cultural knowledge of local, national, and international events. In addition to their evidentiary role, newspapers reflect societal beliefs, reactions, values, and perspectives of current and historical events. Thus, newspapers are a valuable forum for representing how a culture views and reacts to social events. In this study, newspaper stories are one way to reflect not only each state's story of how accountability evolved (e.g., What laws were proposed? How were they debated? When were they passed? How they were implemented?), but also to identify the cultural

norms influencing that state's accountability system (e.g., Was there consensus on each proposal, or were there vehement disagreements?). The inclusion of newspaper articles represents a unique strategy for measuring perceived high-stakes testing pressure.

Scaling

The method of "comparative judgments"[48] was adopted for scaling the study states from low to high according to the relative level of "pressure" associated with its accountability and assessment system. This scaling method was appropriate for assigning relational values among stimuli with complex, abstract psychological properties. Torgerson noted,[49]

> The law of comparative judgment is a set of equations relating the proportion of times any given stimulus *k* is judged greater on a given attribute than another other stimulus *j* to the scale values and discriminal dispersions of the two stimuli on the psychological continuum. The set of equations is derived from the following postulates:
>
> 1. Each stimulus when presented to an observer gives rise to a discriminal process which as some value on the psychological continuum of interest.
> 2. Because of momentary fluctuations in the organism, a given stimulus does not always excite the same discriminal process, but may excite one with a higher or lower value on the continuum. If any stimulus is presented to an observer a large number of times, a frequency distribution of discriminal processes associated with that stimulus will be generated. It is postulated that the values of the discriminal processes are such that the frequency distribution is normal on the psychological continuum.

3. The mean and standard deviation of the distributions associated with a stimulus are taken as its scale value and discriminal dispersions respectively.

The value of this approach is that judges do not have to assign an *absolute* rating to each stimulus. Rather, it is only necessary that judges make a judgment about which of only two stimuli exhibits *more* of the construct of interest. The "stimulus" in this study is the construct of "pressure" as reflected in the portfolio documentation.

Matrix Results

Independent judgments of the pressure associated with each of the 300 possible state pairings were collected. To the judges' data (averaging entries where there were more than one entry per cell), the least-squares solution for uni-dimensional scale values due to Mosteller[50] was used to calculate rating scores. The judges' estimates of the directed distance between any two states on a hypothetical scale of "high stakes pressure" were taken as the raw distance data and formed a skew symmetric matrix of order 25 with entries on the interval -4 to +4 (the results of this conversion are displayed in Tables 7 and 8).

Validity Analysis

As a check on validity of our index, two expert educators also reviewed all 25 portfolios independently rating them on a scale of "pressure" from 1-5. Table 7 displays the results of (a) the PRI results, (b) both experts' rating decisions, (c) both rating systems identified by Carnoy and Loeb, (c) and averaged systems of the experts and of Carnoy and Loeb. Table 8 displays the results of (a) the PRI results, (b) Amrein and

Berliner's initial characterizations, (c) Swanson and Stevenson's policy activism scale, (d) the Boston College classification system, and (e) ECS rating.

Table 7: Comparison of Accountability Measures Across 25 States

	PRI	Expert 1	Expert 2	Carnoy 1	Carnoy 2	Average Expert 1 & 2	Average Expert 1 & Carnoy & Loeb 1	Average Expert 1 & Carnoy & Loeb 2	Average Expert 2 & Carnoy & Loeb 1	Average Expert 2 & Carnoy & Loeb 2
Alabama	3.06	3	2	4	4	2.5	3.50	3.50	3.00	3.00
Arizona	3.36	4.5	4	2	2	4.25	3.25	3.25	3.00	3.00
Arkansas	2.60	2	3	1	1	2.5	1.50	1.50	2.00	2.00
California	2.56	2.5	5	4	2	3.75	3.25	2.25	4.50	3.50
Connecticut	1.60	1.5	1	1	1	1.25	1.25	1.25	1.00	1.00
Georgia	3.44	5.5	4	2	2	4.75	3.75	3.75	3.00	3.00
Hawaii	1.76	0.5	1	1	1	0.75	0.75	0.75	1.00	1.00
Kentucky	0.54	3	3	4	4	2.5	3.50	3.50	3.50	3.50
Louisiana	3.72	5.5	5	3	3	5.25	4.25	4.25	4.00	4.00
Maine	1.78	2	1	1	1	1.5	1.50	1.50	1.00	1.00
Maryland	2.82	2	3	4	5	2.5	3.00	3.50	3.50	4.00
Massachusetts	3.18	4	5	2	2	4.5	3.00	3.00	3.50	3.50
Mississippi	3.82	5.5	2	3	3	3.75	4.25	4.25	2.50	2.50
Missouri	2.14	1.5	3	1.5	1.5	2.25	1.50	1.50	2.25	2.25
New Mexico	3.28	4.5	2	4	5	3.25	4.25	4.75	3.00	3.50
New York	4.08	5.5	5	5	2	5.25	5.25	3.75	5.00	3.50
North Carolina	4.14	3	4	5	5	3.5	4.00	4.00	4.50	4.50
Rhode Island	1.90	1.5	1	1	1	1.25	1.25	1.25	1.00	1.00
South Carolina	3.20	4.5	2	3	3	3.25	3.75	3.75	2.50	2.50
Tennessee	3.50	3	4	1.5	1.5	3.5	2.25	2.25	2.75	2.75
Texas	4.78	5	5	5	5	5	5.00	5.00	5.00	5.00
Utah	2.80	2.5	2	1	1	2.25	1.75	1.75	1.50	1.50
Virginia	3.08	5	4	2	2	4.5	3.50	3.50	3.00	3.00
West Virginia	3.08	1.5	3	3.5	3.5	2.25	2.50	2.50	3.25	3.25
Wyoming	1.00	2	1	1	1	1.5	1.50	1.50	1.00	1.00

Table 8: Comparison of Accountability Measures Across 25 States

	PRI	Amrein & Berliner	Policy Activism	Boston Rating 1*	Boston Rating 2**	ECS
Alabama	3.06	4	2.195	4	9	4
Arizona	3.36	0	-0.395	2	6	2
Arkansas	2.60	0	-0.270	1	8	1
California	2.56	5	0.090	4	9	2
Connecticut	1.60	0	1.290	1	8	1
Georgia	3.44	1	0.660	3	9	3
Hawaii	1.76	0	0.320	1	4	1
Kentucky	0.54	4	1.970	4	7	4
Louisiana	3.72	5	-0.030	3	9	3
Maine	1.78	0	1.290	1	7	1
Maryland	2.82	5	2.460	4	9	5
Massachusetts	3.18	3	0.320	4	9	2
Mississippi	3.82	2	0.550	3	9	3
Missouri	2.14	1	1.020	1	7	1
New Mexico	3.28	5	0.780	4	9	4
New York	4.08	4	0.090	5	9	2
North Carolina	4.14	6	1.600	5	9	5
Rhode Island	1.90	0	0.090	4	7	1
South Carolina	3.20	5	0.900	3	7	3
Tennessee	3.50	4	0.320	3	9	3
Texas	4.78	6	-0.660	5	9	5
Utah	2.80	0	1.150	1	4	1
Virginia	3.08	2	0.550	1	9	1
West Virginia	3.08	3	0.900	3.5	8	3.5
Wyoming	1.00	0	-0.950	1	4	1

* where H/H = 5; H/M or M/H =4; H/L or L/H=3; M/L or L/M=2; and L/L=1
** where H/H=9; H/M=8; H/L=7; M/H=6; M/M=5; M/L=4; L/H=3; L/M=2; L/L=1

Results of a correlation analysis are presented in Tables 9 and 10. Our Pressure Rating Index (PRI) was positively correlated (above .60) with both experts' judgments. Interestingly, the portfolio system, experts' rating judgments, and Carnoy and Loeb's index showed positive, but relatively weak correlations. For example, at one extreme, Expert 2 and Carnoy and Loeb 2 correlated only .29.

Table 9: Correlations of PRI, Experts' and Carnoy and Loeb's Rating Systems

	PRI	Expert 1	Expert 2	Carnoy & Loeb 1	Carnoy & Loeb 2	Average Expert 1 & 2
PRI	1.00					
Expert 1	0.68	1.00				
Expert 2	0.63	0.57	1.00			
Carnoy 1	0.53	0.44	0.51	1.00		
Carnoy 2	0.45	0.34	0.29	0.85	1.00	
Average Expert 1 & 2	0.77	0.89	0.87	0.52	0.34	1.00

In Table 10, among the correlations bearing on the validity of the PRI ratings is the correlation between the newly derived PRI rating and the average of the ratings given by Expert 1 and Carnoy and Loeb 1 (.72), and the correlation of the PRI with the average of Expert 1 and Carnoy and Loeb 2 (.70). The high correlations between some of the other measures (e.g., Amrein & Berliner with either expert averaged with Carnoy & Loeb ratings) most likely resulted from the fact that both Amrein and Berliner and Carnoy and Loeb were essentially counting provisions in the same set of laws.

Table 10: Correlations of PRI, Averaged Ratings, Boston, ECS, and Amrein and Berliner

	PRI	Average Expert 1 & Carnoy & Loeb 1	Average Expert 1 & Carnoy & Loeb 2	Average Expert 2 & Carnoy & Loeb 1	Average Expert 2 & Carnoy & Loeb 2	Amrein & Berliner	Policy Activism	Boston Rating 1*	Boston Rating 2**	ECS
PRI	1.00									
Average Expert 1 & Carnoy & Loeb 1	0.72	1.00								
Average Expert 1 & Carnoy & Loeb 2	0.70	0.95	1.00							
Average Expert 2 & Carnoy & Loeb 1	0.66	0.85	0.75	1.00						
Average Expert 2 & Carnoy & Loeb 2	0.67	0.83	0.83	0.95	1.00					
Amrein & Berliner	0.54	0.75	0.74	0.82	0.85	1.00				
Policy Activism	-0.18	-0.01	0.09	0.00	0.10	0.22	1.00			
Boston Rating 1*	0.51	0.71	0.66	0.77	0.75	0.79	0.14	1.00		
Boston Rating 2**	0.59	0.63	0.62	0.67	0.68	0.64	0.18	0.61	1.00	
ECS	0.49	0.67	0.77	0.67	0.80	0.82	0.38	0.76	0.53	1.00

* where H/H = 5; H/M or M/H =4; H/L or L/H=3; M/L or L/M=2; and L/L=1

** where H/H=9; H/M=8; H/L=7; M/H=6; M/M=5; M/L=4; L/H=3; L/M=2; L/L=1

Because none of the prior measures of high-stakes pressure took into account the actual experience of administrators, teachers, students, and parents subjected to the accountability programs, and because the present empirically-derived PRI index shows consistent positive correlations with indices derived from proxies (features of state laws and regulations) for the actual experience of being subjected to high-stakes testing pressure, the PRI is offered as the most valid measure to date of the construct of "accountability strength" or "high-stakes testing pressure."

Measurement Part II: High-Stakes Pressure Over Time

The PRI represents a judgment of state pressure pooled across all current and past accountability activities made as of summer 2004; therefore, this one-time rating index does not identify when or by how much high-stakes testing pressure grew over the preceding years. For our second set of analyses, we also identified the years during which each state's "pressure" increased and assigned a numerical value to that change. For example, consider a state where a statewide standardized test was first administered to all students in grades 3-8 in 1990. Three years later (1993), the state began holding students back in grades 3 and 8 if they did not pass this test, and in 1999 a law was passed mandating that teachers could be fired or financially compensated based on students' test performance. Given this scenario, it could be argued that prior to 1993 there was "minimal" (if any) pressure on students and teachers to do well on a test. But in 1993, this pressure increased somewhat—most specifically for third and eighth graders and their teachers, and by 1999, the pressure increased again, this time for all teachers. This change in pressure could be depicted the following way:

Year	1990	‘92	‘93	‘94	‘95	‘96	‘97	‘98	‘99
Pressure	1	1	2	2	2	2	2	2	3

Of course, these hypothetical increases are not sensitive to the differential changes in pressure to individual schools, districts, administrators, teachers, and students. Instead, they reflect, as the PRI does, a pooled increase in the amount of pressure as it exists across the entire state.

Assigning values to the timing of accountability implementation was a two-step process. First, one of our education experts read through all 25 portfolios and made a series of judgments about the timing of high-stakes testing increases in each state. On a scale of 0 to 5, this expert assigned a value for the level of threat for each state and for each year from 1985-2004. As a check, a second reader followed the same procedure for a random selection of five portfolios. The results of both readers' judgments on these five states are presented in Table 11.

Table 11: Two Threat Progression Rating Judgments

	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
AR	0	0	0	0	0	0	0	0	0	0	2	2	2	2	3	3	3	4	4	4
AR	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**2**
TN	0	0	0	0	0	0	0	1	1	1	1	1	2	2	2	2	3	3	3	3
TN	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**3**	**3**	**3**	**3**	**3**
MO	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2	2	4	4	4
MO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**
NC	0	0	0	0	0	2	2	2	2	2	2	3	3	3	3	4	4	4	5	5
NC	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**4**	**4**	**4**	**4**	**5**	**5**	**5**	**5**
NM	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3
NM	**0**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**4**	**5**	**5**	**5**	**5**

NOTE: Judgments by Reader 1 are the non-bolded line, and by Reader 2 are the bolded line.

For four out of the five states, both experts agreed on the year during which stakes *first* were attached (e.g., AR, TN, MO, and NM). Also, the experts agreed on two out of the three identified pressure "jump years" for Arkansas (1995, 1999), one out of three for Tennessee (1992), one out of two for Missouri (1993), one out of three for North Carolina (1996), and two out of four for New Mexico (1986 and 2000).

Although experts' judgments did not reach an especially high degree of agreement on the intervening years during which pressure escalated, experts' level of agreement on the year during which stakes were *first* attached to testing was relatively high. Further, experts' level of agreement across the entire time span and the relative amount of "jump" in pressure gain *overall* was relatively consistent (e.g., rating pressure was judged to have doubled for Arkansas and Missouri and viewed to end at the same degree of pressure for Tennessee and North Carolina. But, perhaps more importantly, a second look at Table 9 shows that that Expert 1 had the highest correlation with PRI (r = .68). Expert 2 was only slightly lower in agreement with the PRI (r = .63) and the Carnoy and Loeb indices were well below both experts (r = .53 and .45). Given the impracticality of asking hundreds of judges to rate high-stakes pressure for every year from 1985 to 2003 and for every state, it was decided to let Expert 1 provide all judgments of pressure increase between the years 1985 and 2004 (Table 12). Expert 1 serves as the best available surrogate for the many judges who gave us a robust (albeit cross-sectional) measure of high-sakes testing (i.e., PRI).

Table 12: Finalized Threat Progression Ratings

	'85	'86	'87	'88	'89	'90	'91	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04
Hawaii	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
Rhode Island	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	2	2	2
Missouri	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2	2	2	2
Connecticut	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
West Virginia	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Kentucky*	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
Maryland	0	0	0	0	0	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
Maine	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
Wyoming	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2
Arkansas	0	0	0	0	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	2
Utah	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3
California	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3
Alabama	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
Tennessee	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	3	3	3	3	3
Massachusetts	0	0	0	0	0	0	0	0	1	1	1	1	1	2	2	2	3	3	4	4
South Carolina	0	0	0	0	0	2	2	2	2	2	2	2	2	4	4	4	5	5	5	5
North Carolina*	0	0	0	0	0	0	0	0	0	0	0	3	3	3	3	3	3	3	3	3
New Mexico	0	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	5	5	5	5
Arizona	0	0	0	0	0	2	2	2	2	2	2	2	2	4	4	4	5	5	5	5
Texas*	1	1	1	1	1	3	3	3	3	4	4	4	4	4	5	5	5	5	5	5
Virginia	0	0	0	0	0	0	0	0	0	0	1	1	1	2	2	3	4	4	5	5
Mississippi*	2	2	2	2	2	2	2	2	2	4	4	4	4	4	4	5	6	6	6	6
Louisiana*	1	1	1	1	1	1	4	4	4	4	4	4	4	4	6	6	6	6	6	6
New York	1	1	1	1	2	2	2	2	2	2	2	2	2	2	4	5	6	6	6	6
Georgia	0	0	0	0	0	0	4	4	4	4	4	4	4	4	5	5	5	5	5	6

*These states were evaluated twice. The values here represent the revised judgments.

Methodology

The procedures used to obtain all paired comparison judgments are described in this section. This section starts with a description of the participants who provided their paired comparison judgments followed by the method of analysis used to examine the relationship of pressure and student achievement.

Procedures

The process of comparative judgments requires each individual stimulus (or each state) to be paired with every other. Thus, it was necessary to enlist the participation of 300 individuals who would each review the contents of two state portfolios and provide us with their judgment as to which of the two states exhibited more of the hypothetical construct of "pressure." Graduate-level education students were enlisted for participation because they are familiar with educational terms and activities such as testing and instruction. First, instructors at three major universities in the Southwest were contacted to see if they would be willing to allot two hours of teaching time to this project. Once the help of 13 instructors at the three universities was obtained, a schedule was arranged for times to conduct the study during summer school sessions (May – June, 2004). All 13 instructors provided us with class time. Some offered us the opportunity to come to their class on two occasions. In total, one undergraduate level and 15 graduate-level classes were visited.

All data were collected with groups of students attending summer school courses. Each class was given verbal and written instructions for the task. The following introduction was provided first:

> Thanks for taking the time to participate in this important project. I am working with professors at [university name] on a project to look at the relationship of high-stakes testing across the states with student achievement. I am here asking for your help to determine the relative level of "pressure" associated with the practice of high-stakes testing in each state. More specifically, you will be given two state portfolios and asked to make a single decision: based on the content of both portfolios, which state do you feel exerts a greater level of pressure on the educational players in the state (including students, teachers, and administrators)? You will be making a "pooled" judgment across a wide range of information. For example, you may have one state that exerts pressure on teachers but not students and another state that exerts pressure on students but not teachers. We are asking that you take in all the information and make a generalized assessment about the state, accounting for the pressure to all the players. Simply stated, make your best judgment about which state as a whole exerts more pressure.

Each participant then received a pair of portfolios, a directions sheet (see Appendix C) and an accompanying recording data sheet (see Appendix D). They were then given the following directions:

> You are provided with a data sheet that contains three parts. The first part is a worksheet. Some people like to take notes when they are reading, others do not. It does not matter to me if you fill this out; I have simply provided it for those of

you who like to take notes as they read. The second sheet is the one I am most interested in and is the sheet where you tell me which of your two states has a higher level of pressure than your second state. As you can see, there are two scales. Please assign a single number to each of your states—on a scale of 1 – 7 how much pressure is exerted (where 1 is low pressure and 7 is high pressure). The only rule is that you cannot assign the same number to both states—you must make a decision about which is "higher." You may think one state is higher by one point, or you may determine it is higher by more. Just use your best judgment to make a distinction.

Participants were then given a few last thoughts to help them feel comfortable with the task:

Everyone reads at a different rate, some may be able to go through the portfolios and make a decision in about an hour; others may take 2 or more hours. Please take as much time as you need to make the best judgment possible. Note that there is highlighting throughout the portfolio. This was done not to dissuade you from reading, but to facilitate your reading. Please read as much as you need again, to make the best judgment you can.

Also, there are no right or wrong answers. I have no extant expectations on how each of these states compare, you are helping me figure that part out. Lastly, although no one has the same TWO states, many of you may have one state that is the same. For example, student A may have Arizona and Maine, and student B may have Arizona and Texas. It is perfectly acceptable to talk about the

contents of your portfolios. I only ask that you do not discuss your pending judgments as everyone is making a judgment based on the two states they have.

It was also pointed out that throughout the portfolio significant portions were highlighted. This was done to facilitate the process of reviewing so much information in one sitting. Anecdotal information suggested that highlighting sections of the portfolio greatly facilitated the process without significantly altering judgments about the portfolio as a whole.[51]

Participants

A total of 346 paired comparison judgments were collected. The number of individuals who provided the judgments was fewer than 346 since several individuals participated more than once (in one case, three times). It is difficult to accurately assess the number of participants since all data was collected anonymously. However, judgments from approximately 250 different persons were obtained. Of the total 346 paired comparisons, 239 (69 percent) were provided by females and 93 (27 percent) by males, with gender missing on 14 (4 percent). Many participants had taught for some period of time in a K-12 or university setting. There were 254 (73 percent) participants who replied "yes" and 77 (22 percent) who replied "no" to the question, "Have you ever taught?" (Fourteen provided no data.) Most participants were in a graduate school program with 313 specifying they were in one of the following degree programs: MA (142), EdD (22), PhD (32), or graduate level school, degree unspecified (117). There were only 14 who were in an undergraduate program and one participant in a post-baccalaureate program.

Feedback on Method

Time for data collection varied from one to three hours per person comparing two states. After every data collection session, some participants were asked to provide feedback on their confidence level for their judgments. An overwhelming majority of those asked felt confident that they made the right judgment. Further, participants often reported that (a) the task was very interesting and (b) that at least one of their states stood out as having more pressure than the other. Comments also included that the "task was interesting," that "they couldn't believe the dramatic differences between states," or that "they had no idea how bad some states had it." For those who were teaching at the time of the task, many felt relieved they did not live in another state they perceived to be dramatically greater in the pressure exerted on teachers and students than what they were experiencing. Many noted, "Thank goodness I don't work in state X," or, "I will never move to state X."

Participants were also asked their strategy for using the materials provided to them. It was clear that strategies varied widely. Some participants relied heavily on the rewards/sanctions worksheet whereas others thought the newspaper documents helped them more. Some used the comparison sheets as a starting point and went back and forth between portfolios on each specific document, whereas others would go through one portfolio before looking at the second one.

Method of Analysis

Four approaches were used in our analyses. First, we used our accountability rating system to replicate Carnoy and Loeb's analysis and to test their conclusion that

high-stakes testing is related to achievement gains for minority students. This included the replication of three regression models.

Carnoy and Loeb's first regression model estimates accountability implementation as a function of the average level of National Assessment of Education Progress (NAEP) test scores in each state in the early 1990s, test score gains in the early 1990s, the percent of Latinos and African Americans in the state, the state population, the percent of school revenues raised at the state level in 1995,[52] average per-pupil revenues in 1990, and the yearly change in revenues in the early 1990s:

(1) $A_i = \beta_0 + \beta_1 T_i + \beta_2 R_i + \beta_3 P_i + \beta_4 S_i + \beta_5 D_i$ €

Where,

A = strength of accountability in state (measured by our rating system);

T = average scale score of fourth grade students in state on the 1992 math NAEP;

R = the proportion of African American and Hispanic (public school) students in state *i*;

P = the state population; and

S = the proportion of schools' funds coming from the state rather than local sources in 1995; and

D = Dollars per pupil revenues in 1990 and the yearly percent change in revenue from 1990 to 1995.

Carnoy and Loeb's second regression tests whether the proportion of eighth graders (or fourth graders) achieving at the basic skills level or better (and at the

proficient level or better on the NAEP math test) increased more between 1996 and 2000 in states with "stronger" accountability.[53] Again, we adopted their regression equation:

(2) $G_i = \phi_0 + \phi_1 A_i + \phi_2 M_i + \phi_3 T_i \text{ (or } H_i) + \phi_4 S_i + \epsilon$

Where,

G = the change in the proportion of eighth grade students in state *i* who demonstrate basic skills or better on the mathematics NAEP between 1996 and 2000;

A = strength of accountability in state (measured by our PRI system);

M = the proportion of African American and Hispanic (public school) students in state *i*;

T = the average percentage of eighth grade students in state *i* demonstrating basic math skills or better or demonstrating proficient level or better on the mathematics NAEP in 1996;

H = the change in the average percentage of eighth grade students in state *i* demonstrating basic math skills or better on the mathematics NAEP between 1992 and 1996;

S = a dichotomous variable indicating whether state is in the South.

In terms of their third regression model, we looked at whether ninth-grade retention rates increased more in the late '90s in states with "strong" accountability than in states with "weak accountability."

(3) $Rt_i \text{ or } Pg_i = \Theta_0 + \Theta_1 A_i + \Theta_2 T_i + \Theta_3 M_i + \Theta_4 P_i + \Theta_5 S_i + \epsilon$

Where,

Rt = the ninth grade retention rate in state *i*;

Pg = the high school progression rate in state *i*; and

T = NAEP eighth grade math test scores in 1996.

The second part of our results includes a series of correlations investigating the relationship between overall changes in high-stakes testing "pressure" and overall achievement gains. First, we analyze whether pressure is associated with achievement gains between the very first year of NAEP administration and the most recent. Then the relationship between change in pressure rating and NAEP gains by student cohort is analyzed. Lastly, a series of correlations investigating whether *prior* changes in high-stakes testing pressure is related to subsequent changes in NAEP achievement (both in terms of a cross-sectional and cohort strategy) is calculated.

Data

Data from NAEP tests were used as the achievement indicator for fourth- and eighth-grade math and reading.[54] NAEP data included both scale score and proficiency percentages at the state level and disaggregated by ethnicity. Demographic information for the Carnoy and Loeb replication analysis including percent of African American and Hispanic students in each state as of 1995, percent of school funds coming from state rather than local revenues,[55] and state population demographic characteristics[56] were drawn from a variety of internet sources.

Results

Part I: Carnoy and Loeb Replication

Carnoy and Loeb conducted a series of analyses to test the relationship of their strength of accountability index against a range of achievement and demographic variables. Their analyses are replicated, substituting our PRI for their index.[57]

Replicating Carnoy and Loeb's Equation One

To test whether our accountability measure was related to various demographic variables identified by Carnoy and Loeb, correlation and regression analyses were computed (see Tables 13 and 14).

Table 13: Correlations of PRI, Demographic, and Achievement Variables: NAEP Fourth-Grade Math and Reading

	A	B	C	D	E	F	G	H	I	J	K	L
A: PRI	1.000											
B: State Population (1995), used 1990 Census Data	0.357	1.000										
C: Proportion African American and Hispanic in state in 1995	0.675	0.519	1.000									
D: 1992 math fourth-grade White scale score	-0.114	0.035	-0.097	1.000								
E: 1992 math fourth grade African American scale score	0.364	0.237	0.526	-0.109	1.000							
F: 1992 fourth-grade reading White, % at least basic	-0.109	-0.004	-0.164	0.657	-0.306	1.000						
G: 1992 fourth-grade reading African American, % at least basic	0.106	-0.093	-0.135	-0.341	-0.136	-0.255	1.000					
H: Change in fourth-grade reading 1992-1994 White	0.045	-0.292	-0.144	-0.115	0.105	-0.001	0.141	1.000				
I: Change in fourth-grade reading 1992-1994 African American	-0.403	-0.208	-0.473	-0.125	-0.353	-0.032	0.344	0.225	1.000			
J: Yearly percent revenue change 1990-1995 (Average of all yearly changes)	0.019	-0.274	-0.210	-0.327	-0.053	-0.396	0.482	0.111	0.429	1.000		
K: Proportion of revenues coming from state (not local or federal)	-0.146	-0.242	-0.161	-0.501	0.046	-0.641	0.495	0.117	0.252	0.312	1.000	

L: Average Per pupil Revenue 1990-1991	-0.191	0.143	-0.126	0.308	0.077	0.584	-0.221	0.248	-0.056	-0.554	-0.411	1.000

These correlations suggest that corresponding to what Carnoy and Loeb found, state composition (those with a higher proportion of African American and Hispanic students) is related to accountability pressure. However, in contrast to their finding that states with lower NAEP scores in the early 1990s implemented stronger accountability systems later (for White students), there is no evidence that pressure is associated with early NAEP performance among Whites. Instead, pressure is positively related to math fourth-grade African American scale score (1992) but negatively correlated to the change in fourth-grade reading scale scores (1992-1994) for African American students.

Our regression model was not significant (See Table 14). When all demographic and achievement variables were entered simultaneously, the only significant predictor of accountability pressure was the state composition variable—states with a greater proportion of minority students (African American and Hispanic) implemented accountability systems that fostered greater pressure.

Table 14: Regression Model: Predicting Accountability From Achievement and Demographic Variables

ANOVA					
	df	**SS**	**MS**	**F**	**Significance F**
Regression	11.000	15.285	1.390	1.980	0.121
Residual	13.000	9.122	0.702		
Total	24.000	24.407			

Table 14, continued

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	2.478	5.935	0.417	0.683	-10.344	15.300	-10.344	15.300
1995 Census Estimates	0.000	0.000	0.702	0.495	0.000	0.000	0.000	0.000
Proportion African American and Hispanic 1995	2.866	1.785	1.606	0.132	-0.989	6.721	-0.989	6.721
1992 fourth-grade math White scale score	-0.009	0.026	-0.366	0.720	-0.065	0.047	-0.065	0.047
1992 fourth-grade math African American scale score	0.001	0.004	0.356	0.727	-0.007	0.010	-0.007	0.010
1992 fourth-grade reading White, % at least basic	3.321	6.098	0.545	0.595	-9.854	16.496	-9.854	16.496
1992 fourth-grade reading African American, % at least basic	2.430	2.046	1.188	0.256	-1.990	6.850	-1.990	6.850
Change in fourth-grade reading 1992-1994 White	0.091	0.071	1.271	0.226	-0.063	0.245	-0.063	0.245
Change in fourth-grade reading 1992-1994 African American	-0.048	0.041	-1.189	0.256	-0.136	0.040	-0.136	0.040
Yearly percent revenue change 1990-1995 (Average of all yearly changes)	3.990	25.160	0.159	0.876	-50.365	58.345	-50.365	58.345
Proportion of revenues coming from state (not local or federal)	-1.599	2.153	-0.742	0.471	-6.250	3.053	-6.250	3.053
Average per pupil Revenue 1990-1991	0.000	0.000	-1.036	0.319	-0.001	0.000	-0.001	0.000

Table 14, continued

Regression Statistics	
Multiple R	0.791
R Square	0.626
Adjusted R Square	0.310
Standard Error	0.838
Observations	25.000

Replication of Carnoy and Loeb's Equation Two

Carnoy and Loeb's second regression model is replicated to test whether accountability was related to achievement. In this set of analyses, Carnoy and Loeb included a measure of whether a state was located in the south—a variable identified by others.[58] Importantly, Carnoy and Loeb's definition of what state was in the south was unclear; therefore, our findings are presented based on all possible characterizations. Correlations and regression results for eighth-grade achievement are presented in Tables 15 and 16, and correlations and regression results for fourth-grade achievement are presented in Tables 19 and 20.

Table 15: Correlations of PRI With Proportion of Students Achieving at Basic or Better: Eighth-Grade Math (1996-2000)

	A	B	C	D	E	F	G	H	I
A: PRI	1.000								
B: Change in % at or above basic eighth-grade math 1996-2000	0.446	1.000							
C: Proportion African American and Hispanic 1995	0.676	0.394	1.000						
D: Eighth-grade math at or above basic 1996	-0.404	-0.446	-0.591	1.000					
E: Eighth-grade math at or above proficient 1996	-0.301	-0.306	-0.408	0.937	1.000				
F: Change in % at or above basic eighth-grade math 1992-1996	0.092	0.045	-0.253	0.124	0.111	1.000			
G: State in South? (0=no; 1=yes)*	0.466	0.475	0.426	-0.644	-0.686	0.158	1.000		
H: State in South? (0=no; 1=yes)**	0.387	0.613	0.274	-0.599	-0.614	0.258	0.852	1.000	
I: State in South? (0=no; 1=yes)***	0.232	0.511	0.153	-0.624	-0.649	0.189	0.786	0.923	1.000

* with southwestern states (AZ, NM, and TX as yes)
** with AZ and NM as no, and TX as yes
*** with AZ, NM, and TX as no

These correlations reveal a positive relationship between pressure, as measured by the PRI, and the change in the percent of students at or above basic in eighth-grade math later in the 1990s (1996-2000). However, we wondered if this positive correlation was confounded by increases in exclusion rates; therefore we calculated a partial correlation holding 2000 NAEP exclusion rates constant. For this (and all subsequent partial correlation equations) we adopt the equation:

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

where:

r_{12} = Correlation of NAEP indicator and PRI indicator;

r_{13} = Correlation of NAEP indicator and exclusion rate;

r_{23} = Correlation of PRI indicator and exclusion rate.

When exclusion rates are partialed out of the relationship, the correlation drops to essentially zero ($r = .026$).

A regression analysis that assesses whether pressure (PRI) or any demographic variables predict changes in the percent of students at or above basic in eighth-grade math between 1996-2000 is significant and is largely explained by a negative effect of yearly percent change in state-revenue (1990-1995) and not high-stakes testing pressure.

Table 16: Regression Model: Predicting Eighth-Grade Math NAEP Change (1996-2000) From PRI and Demographic Variables.

ANOVA					
	df	SS	MS	F	Significance F
Regression	12.000	362.223	30.185	3.378	0.022
Residual	12.000	107.217	8.935		
Total	24.000	469.440			

Table 16, continued

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	22.254	14.053	1.584	0.139	-8.364	52.873	-8.364	52.873
PRI	1.277	0.938	1.362	0.198	-0.766	3.321	-0.766	3.321
1996 eighth-grade math at or above basic	-0.343	0.165	-2.079	0.060	-0.703	0.016	-0.703	0.016
Proportion African American and Hispanic 1995	-16.769	10.861	-1.544	0.149	-40.433	6.894	-40.433	6.894
1995 Census Estimates	0.000	0.000	-0.496	0.629	0.000	0.000	0.000	0.000
Proportion of revenues coming from state (not local or federal)	-6.402	6.833	-0.937	0.367	-21.291	8.486	-21.291	8.486
Average Per pupil Revenue 1990-1991	0.000	0.001	0.617	0.549	-0.001	0.002	-0.001	0.002
Yearly percent revenue change 1990-1995 (Average of all yearly changes)	-203.730	80.830	-2.520	0.027	-379.844	-27.617	-379.844	-27.617
Change in Population 1996-2000	17.690	38.672	0.457	0.656	-66.569	101.949	-66.569	101.949
Change proportion of African American/Hispanic Students 1996-2000	30.129	12.331	2.443	0.031	3.262	56.996	3.262	56.996
State in South?*	-2.837	3.435	-0.826	0.425	-10.321	4.647	-10.321	4.647
State in South?**	13.249	4.664	2.841	0.015	3.088	23.410	3.088	23.410
State in South?***	-6.214	4.204	-1.478	0.165	-15.374	2.947	-15.374	2.947

* with southwestern states (AZ, NM, and TX as yes)
** with AZ and NM as no, and TX as yes
*** with AZ, NM, and TX as no

Table 16, continued

Regression Statistics	
Multiple R	0.878
R Square	0.772
Adjusted R Square	0.543
Standard Error	2.989
Observations	25.000

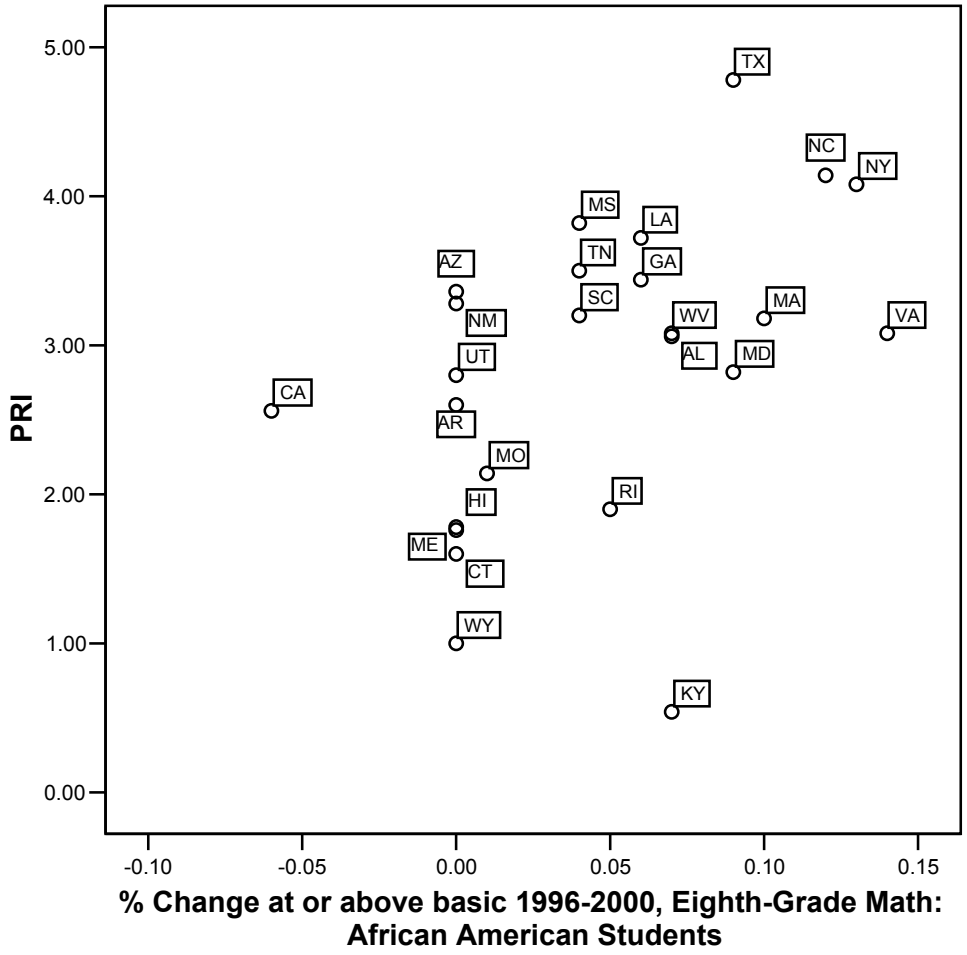
Similar to analyses by Carnoy and Loeb, another set of analyses was done by disaggregating the data by student ethnicity. Correlation results (Table 17) suggest that pressure is associated with changes in the percentages of students who achieve at basic or above (again, eighth-grade math, 1996-2000) for African American students but not for White or Hispanic students (correlations between achievement indicators and PRI are in bold).

Table 17: Correlation of Eighth-Grade Math NAEP Performance with Demographic Variables and PRI and Disaggregated by Student Ethnicity

	A	B	C	D	E	F	G	H	I
A: PRI	1.000								
B: Proportion African American and Hispanic 1995	0.675	1.000							
C: 1995 Census Estimates	0.357	0.519	1.000						
D: Average per pupil revenue 1990-1991	-0.191	-0.126	0.143	1.000					
E: Change in population 1996-2000	0.429	0.429	0.281	-0.390	1.000				
F: Change proportion of African American/Hispanic students 1996-2000	-0.046	0.168	0.094	0.133	0.113	1.000			
G: Change % at or above basic 1996-2000 Hispanic	**0.094**	0.055	0.270	0.325	-0.495	-0.086	1.000		
H: Change % at or above vasic 1996-2000 African American	**0.456**	0.170	0.036	0.272	-0.032	0.021	0.384	1.000	
I: Change % at or above basic 1996-2000 White	**0.054**	-0.078	0.016	0.242	0.201	0.135	-0.017	0.419	1.000

A scatter plot of the relationship between change in percent at or above basic (1996-2000) for African American students and PRI suggests there are no conspicuous outliers (see Figure 1).

Figure 1: Scatter Plot of Change in Percent of Eighth-Grade Students At or Above Basic (1996-2000) NAEP Math and PRI: African American Students



Correlation analyses substituting NAEP scale scores for percent scoring at or above basic were calculated (see Table 18). The relationship between average NAEP scale score gains from 1996-2000 and pressure is low but positive for students overall and when disaggregated by student ethnicity.

Table 18: Correlation of Eighth-Grade Math NAEP Average Scale Score Gains Disaggregated by Ethnicity, PRI, and State Demographic Variables

	A	B	C	D	E	F	G	H	I	J
A: PRI	1.000									
B: Proportion African American and Hispanic 1995	0.675	1.000								
C: 1995 Census Estimates	0.357	0.519	1.000							
D: Average Per Pupil Revenue 1990-1991	-0.191	-0.126	0.143	1.000						
E: Change in Population 1996-2000	0.429	0.429	0.281	-0.390	1.000					
F: Change proportion of African American/ Hispanic Students 1996-2000	-0.046	0.168	0.094	0.133	0.113	1.000				
G: NAEP Gain 1996-2000 All	**0.372***	0.227	0.044	0.234	-0.009	0.211	1.000			
H: NAEP Gain 1996-2000 White	**0.213**	0.112	-0.068	0.284	0.015	0.272	0.872	1.000		
I: NAEP Gain 1996-2000 African American	**0.274**	0.119	0.056	0.396	-0.259	0.019	0.715	0.512	1.000	
J: NAEP Gain 1996-2000 Hispanic	**0.314**	0.228	0.370	0.566	-0.116	0.057	0.358	0.277	0.322	1.000

* Partial correlation holding 2000 exclusion rates constant is .320.

Scatter plots of all NAEP scale and gain scores with PRI are presented in Figures 2, 3, and 4. For white students, a correlation between NAEP gain and PRI eliminating North Carolina as a potential outlier (with NAEP gain of 9) lowers the overall relationship to r =.085. There are no conspicuous outliers for African American or Hispanic students.

Figure 2: Scatter Plot of Eighth-Grade White Students NAEP Scale Score Gain and PRI

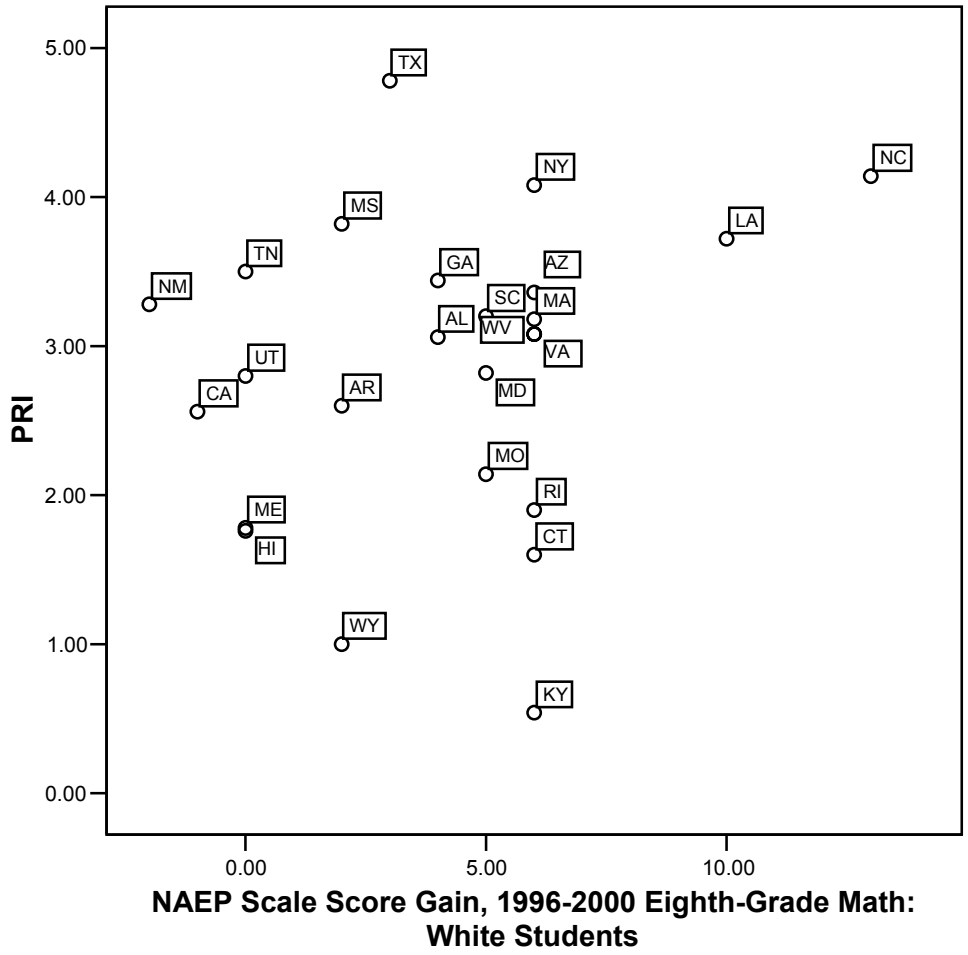


Figure 3: Scatter Plot of Eighth-Grade African American Students NAEP Scale Score Gain and PRI

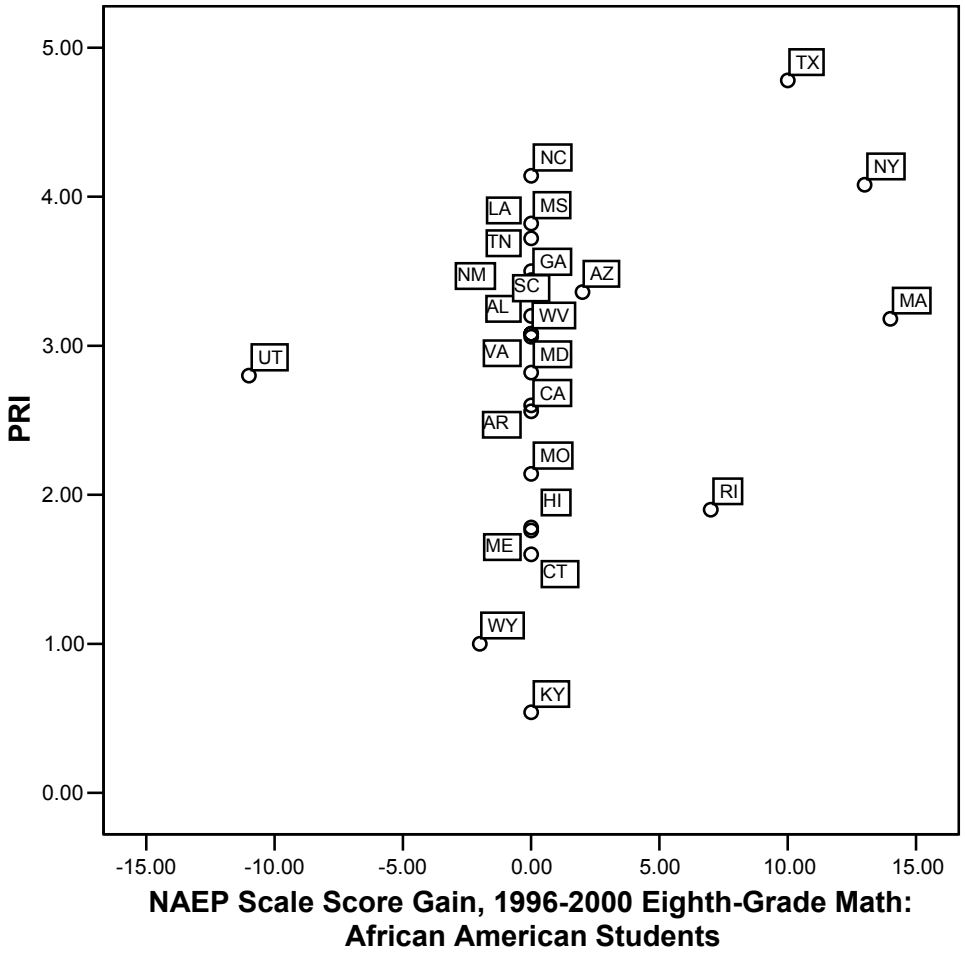
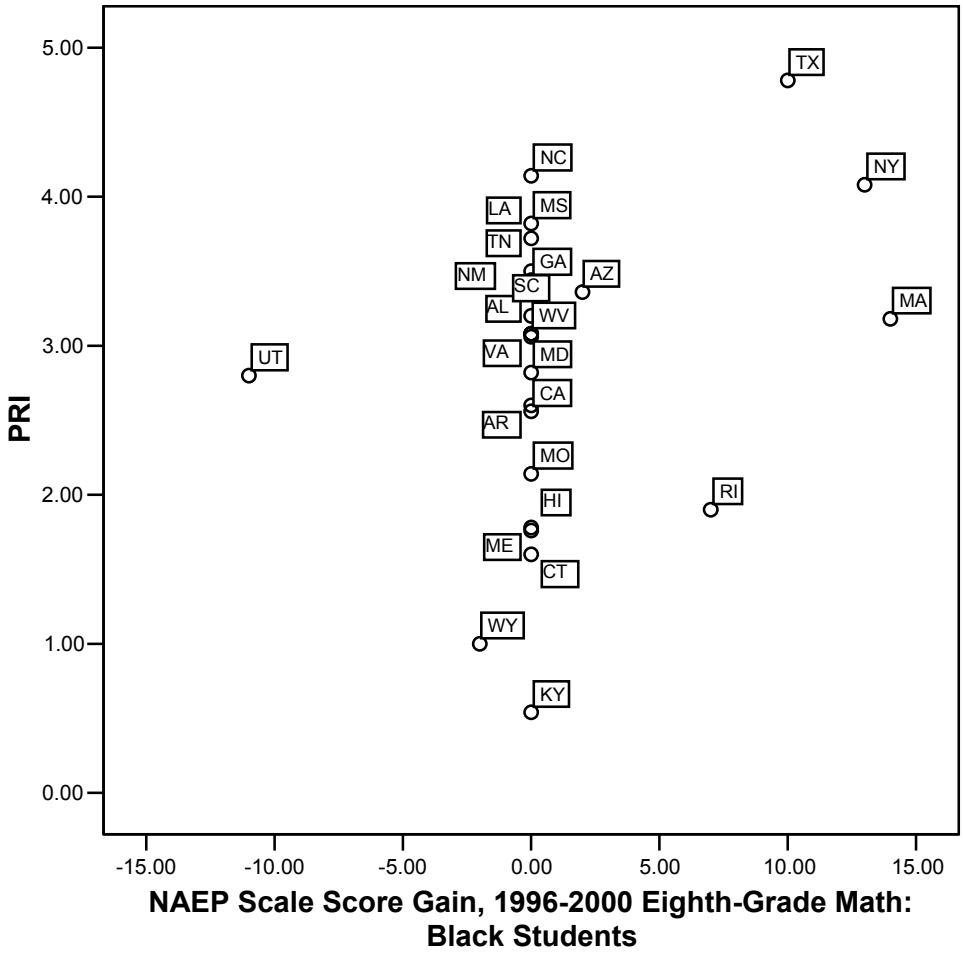


Figure 4: Scatter Plot of Eighth-Grade Hispanic Students NAEP Scale Score Gain and PRI



The same correlations and regression models are calculated for fourth-grade math data. First, a series of correlations looking at the relationship between pressure and change in percent of students achieving at basic and/or proficiency or above during 1996-2000 was calculated. A positive relationship between overall pressure and a change in the percentage of students achieving at basic or above from 1996-2000 (Table 19) was found.

Table 19: Correlations of Fourth-Grade NAEP Achievement, PRI, and Demographic Variables

	A	B	C	D	E	F	G	H
A: PRI	1.000							
B: Change in percent at or above basic 1996-2000 All	**0.350**	1.000						
C: Proportion African American and Hispanic 1995	0.675	0.378	1.000					
D: Percent at or above basic, fourth-grade math 1996 All	-0.227	-0.270	-0.552	1.000				
E: Percent at or above proficient, fourth-grade math 1996 All	-0.180	-0.268	-0.439	0.960	1.000			
F: State in South?*	0.466	0.245	0.426	-0.471	-0.512	1.000		
G: State in South?**	0.387	0.420	0.274	-0.380	-0.408	0.852	1.000	
H: State in South?***	0.232	0.356	0.153	-0.470	-0.515	0.786	0.923	1.000

* with southwestern states (AZ, NM, and TX as yes)
** with AZ and NM as no, and TX as yes
*** with AZ, NM, and TX as no
NOTE: Partial correlation of PRI and change in percent at or above basic, 1996-2000 holding 2000 NAEP exclusion rates constant is .346

We regressed our pressure index along with demographic and achievement variables against the change in percent of students achieving at basic or above from 1996-2000 for

fourth-grade math. Our regression was significant and was largely explained by year percent revenue change (1990-1995) and not pressure associated with high-stakes testing (see Table 20).

Table 20: Regression Model: Predicting Changes in NAEP Proficiency—Fourth-Grade Math

ANOVA					
	df	**SS**	**MS**	**F**	**Significance F**
Regression	13.000	0.030	0.002	3.500	0.022
Residual	11.000	0.007	0.001		
Total	24.000	0.038			

Table 20, continued

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	0.027	0.163	0.168	0.870	-0.330	0.385	-0.330	0.385
PRI	0.014	0.009	1.587	0.141	-0.005	0.032	-0.005	0.032
At or above basic 1996 All	0.434	0.345	1.259	0.234	-0.325	1.194	-0.325	1.194
Proportion African American and Hispanic 1995	0.072	0.114	0.631	0.541	-0.179	0.322	-0.179	0.322
1995 Census Estimates	0.000	0.000	-1.866	0.089	0.000	0.000	0.000	0.000
Proportion of 1995 revenues coming from state (not local or federal)	-0.039	0.063	-0.612	0.553	-0.178	0.101	-0.178	0.101
Average er pupil revenue 1990-1991	0.000	0.000	-1.744	0.109	0.000	0.000	0.000	0.000
At or above proficient 1996 fourth-grade math	-0.899	0.473	-1.899	0.084	-1.940	0.143	-1.940	0.143
Yearly percent revenue change 1990-1995 (average of all yearly changes)	-2.857	0.698	-4.094	0.002	-4.393	-1.321	-4.393	-1.321
Change in population 1996-2000	-0.453	0.328	-1.383	0.194	-1.175	0.268	-1.175	0.268
Change proportion of African American / Hispanic students 1996-2000	0.204	0.111	1.844	0.092	-0.040	0.448	-0.040	0.448
State in South?*	-0.102	0.035	-2.894	0.015	-0.180	-0.025	-0.180	-0.025
State in South?**	0.155	0.047	3.289	0.007	0.051	0.259	0.051	0.259
State in South?***	-0.054	0.040	-1.345	0.206	-0.144	0.035	-0.144	0.035

* with southwestern states (AZ, NM, and TX as yes)
** with AZ and NM as no, and TX as yes
*** with AZ, NM, and TX as no

Table 20, continued

Regression Statistics	
Multiple R	0.897
R Square	0.805
Adjusted R Square	0.575
Standard Error	0.026
Observations	25.000

The same set of analyses for fourth-grade math was calculated based on data disaggregated by ethnicity. Table 21 displays all correlations among PRI, demographic variables, and fourth-grade achievement indicators.

Table 21: Correlations of Fourth-Grade Math Changes in Percent Proficiency (1996-2000) and Disaggregated by Ethnicity

	A	B	C	D	E	F	G	H	I
A: PRI	1.000								
B: Proportion African American and Hispanic 1995	0.675	1.000							
C: 1995 Census Estimates	0.357	0.519	1.000						
D: Average per pupil revenue 1990-1991	-0.191	-0.126	0.143	1.000					
E: Proportion of 1995 revenues coming from state (not local or federal)	-0.146	-0.161	-0.242	-0.411	1.000				
F: Change in proportion of African American/Hispanic students 1996-2000	-0.046	0.168	0.094	0.133	-0.071	1.000			
G: Change in percent at or above basic 1996-2000 White	**0.184**	0.322	0.134	-0.013	-0.340	0.042	1.000		
H: Change in percent at or above basic 1996-2000 Hispanic	**0.281**	0.497	0.187	-0.046	0.055	0.141	0.512	1.000	
I: Change in percent at or above basic 1996-2000 African American	**0.327**	0.117	0.273	0.105	-0.467	0.075	0.430	0.167	1.000

All correlations are positive but relatively weak for change in percent scoring at basic and above and PRI for White and Hispanic students. But the correlation for African American students is somewhat stronger.

Two scatter plots (Figures 5 and 6) of the relationship between PRI and changes in percent basic among Hispanic and African American students reveal two significant outliers. Correlations were computed eliminating these outliers changing the correlation between PRI and percent at or above basic among Hispanic students to .196 (after eliminating Maine) and to .713 among African American students after eliminating New Mexico.

Figure 5: Scatter Plot of Change in the Percent of Fourth-Grade Students At or Above Basic and PRI: African American Students

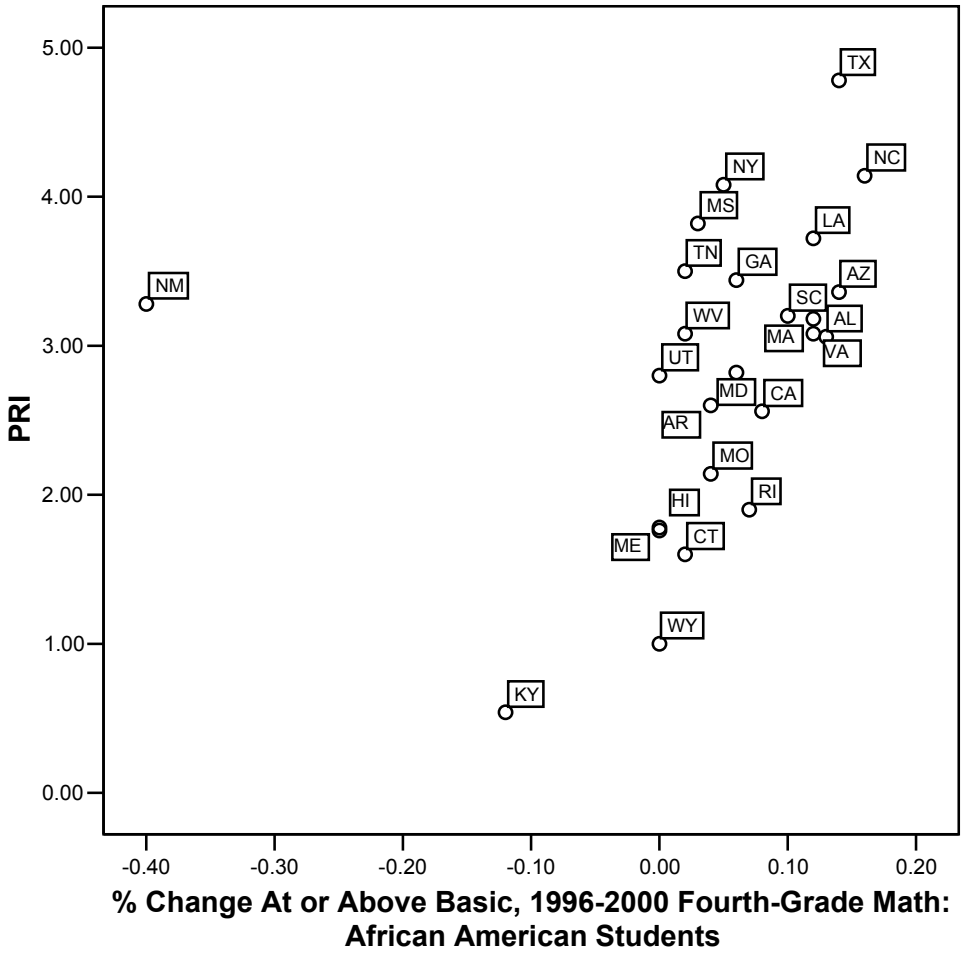
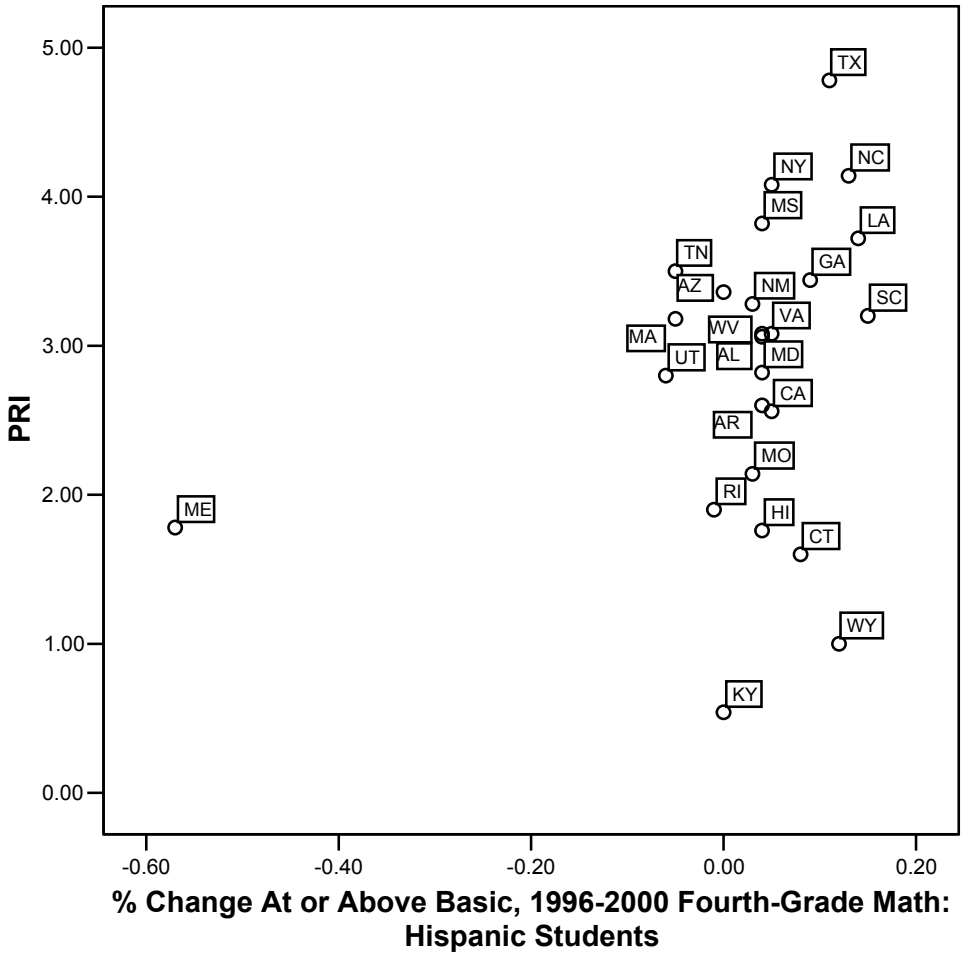


Figure 6: Scatter Plot of Change in the Percent of Fourth-Grade Students At Basic or Above and PRI: Hispanic Students



Replication of Carnoy and Loeb's Equation Three

We did not have the exact estimates of retention and progression rates calculated by Carnoy and Loeb. However, we adopted their procedures for calculating progression. Using enrollment data[59] we estimated progression in terms of (a) the ratio of the number of students in ninth grade in year *i* related to the number in eighth grade in year *i* -1 for the ninth-grade progression rate, (b) the ratio of the number of students in 12th grade in year *i* related to the number in 10th grade in year *i* -2 for the 10th-12th grade progression rate, and (c) the ratio of the number of 12th grade students in year *i* related to the number of students in eighth-grade in year *i*-4 for the high school progression rate.

As shown in Table 22, high-stakes testing pressure is positively correlated with the probability that students progress from eighth- to ninth-grade. Interestingly, in spite of relatively strong correlations in prior years, the relationship for the most recent years where enrollment data is available is relatively weak (i.e., 2000-2001). By stark contrast, the relationships between PRI and eighth- and 10th-grade progression into 12th-grade were all negative.

Table 22: Correlation of Eighth-Ninth-Grade Progression Rates and PRI

8th-9th		10th-12th		8th-12th	
1993-1994	.424	1993-1995	-.513	1993-1997	-.434
1994-1995	.499	1994-1996	-.438	1994-1998	-.442
1995-1996	.446	1995-1997	-.443	1995-1999	-.411
1996-1997	.462	1996-1998	-.401	1996-2000	-.353
1997-1998	.365	1998-2000	-.342	1997-2001	-.386
1998-1999	.416	1999-2001	-.331		
1999-2000	.415				
2000-2001	.188				

Part II: Relationship of Change in PRI and Change in NAEP Achievement

We conducted a series of correlations and partial correlations to examine the relationship between NAEP gains (between first administration and the most recent one) first administration and the most recent one) and pressure rating index change (across the same years). Table 23 displays all correlations by grade level and disaggregated by ethnicity for fourth- and eighth-grade math and reading (A complete set of scatter plots, correlations and partial correlations for all years, grades, and subject areas aggregated at the state level and disaggregated by student ethnicity is available in Appendix E).

Table 23: Correlations and Partial Correlations of NAEP Gain and Threat Rating Change

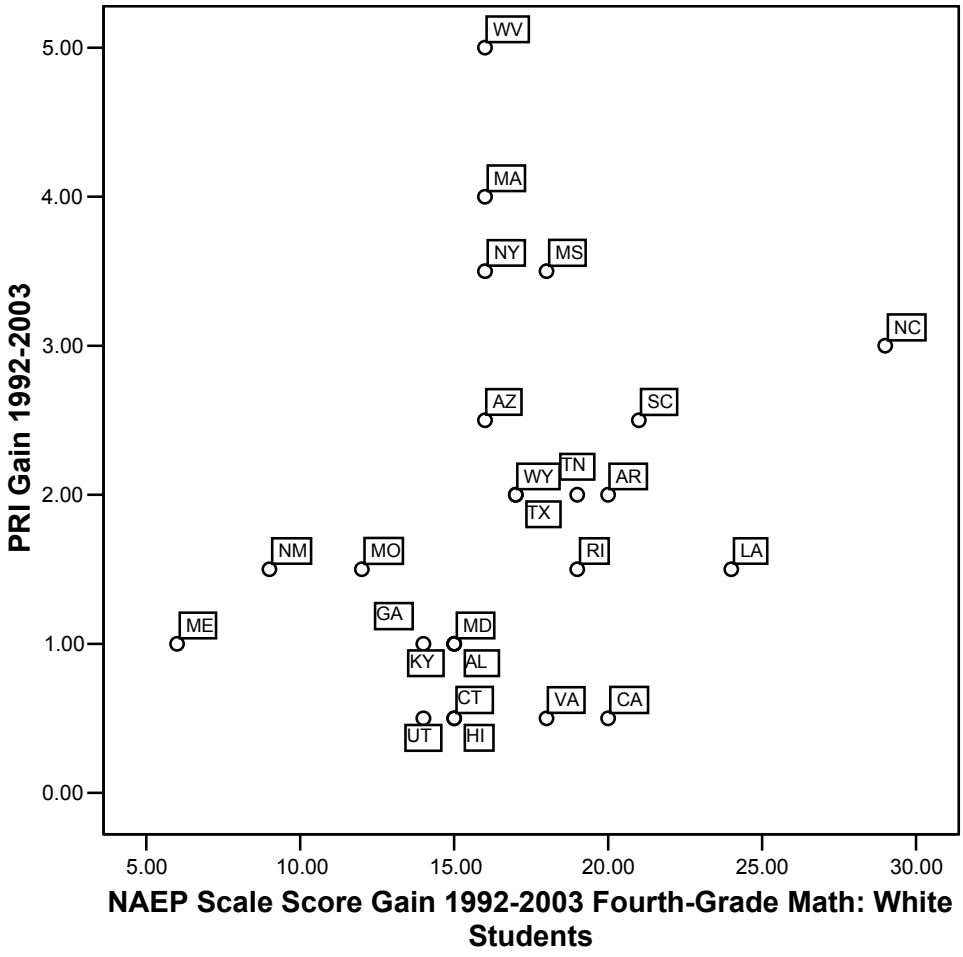
	MATH		READING	
	Grade 4*	**Grade 8****	**Grade 4***	**Grade 8*****
***r* for all students**	.370	.283	.187	.170
Partial *r* for all students	.343	.280	.157	.198
***r* for African American students**	.194	.330	-.060	.109
Partial *r* for African American students	.161	.315	-.077	.081
***r* for Hispanic students**	.383	.112	-.007	.243
Partial *r* for Hispanic students	.370	.077	.024	.251
***r* for White students**	.254	-.106	.159	.264
Partial *r* for White students	.244	-.098	.136	.217

Note: Partial *r* is same correlation holding 2003 NAEP exclusion rates constant.
* Based on NAEP gain scores and threat rating change calculated as 2003 data – 1992 data.
**Based on NAEP gain scores and threat rating change calculated as 2003 data – 1990 data.
***Based on NAEP gain scores and threat rating change calculated as 2003 data – 1998 data.

Fourth-Grade Math

Looking at the change between 1992 and 2003 and aggregated across all students, the relationship between NAEP gain and the increase in high-stakes testing pressure is moderately high; however, when the data are disaggregated by ethnicity, it can be seen that this relationship is primarily explained by Hispanic and White student gains. The relationship between pressure and achievement change among African American students is weak. A series of scatter plots reveals that among White students, a potential outlier may be influencing the strength of the correlation (see Figure 7). A correlation eliminating North Carolina (with a NAEP 1992-2003 gain of 29 points) yielded an even lower relationship (.174) between NAEP gain and threat change among White students.

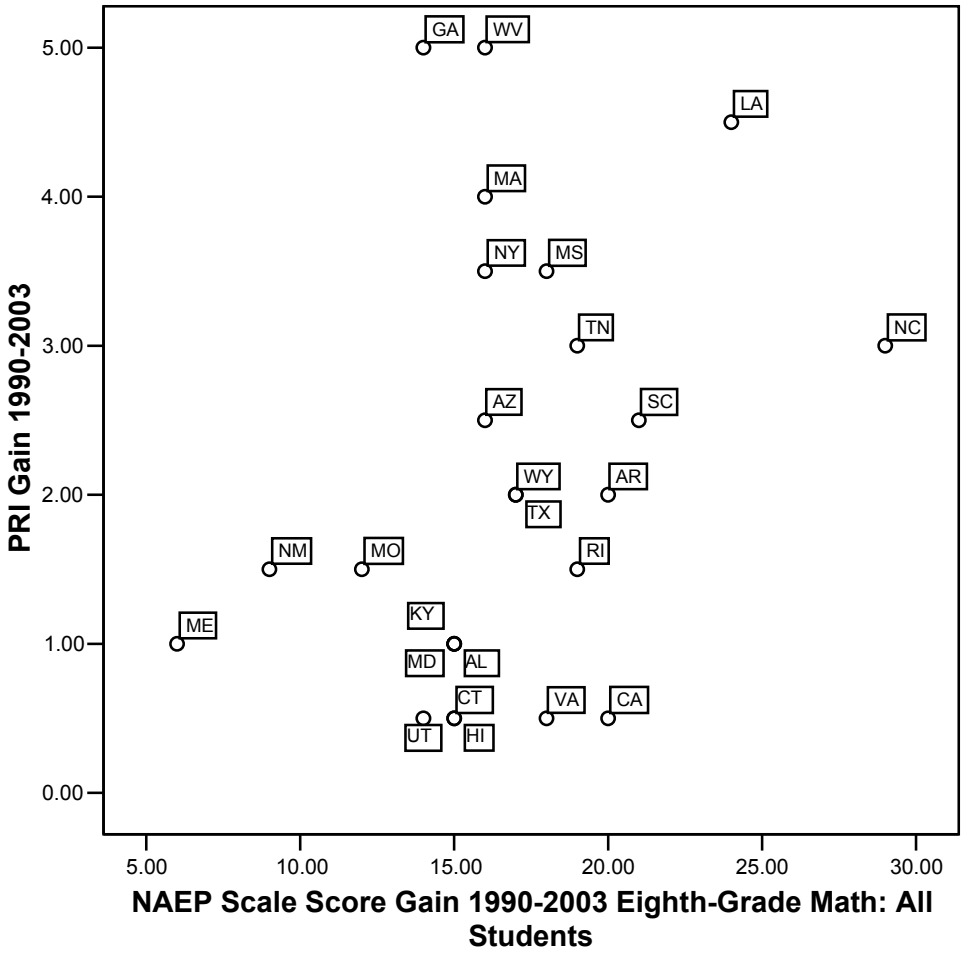
Figure7: Scatter Plot of Fourth-Grade Math NAEP Gain and Threat Rating Change: White Students



Eighth-Grade Math

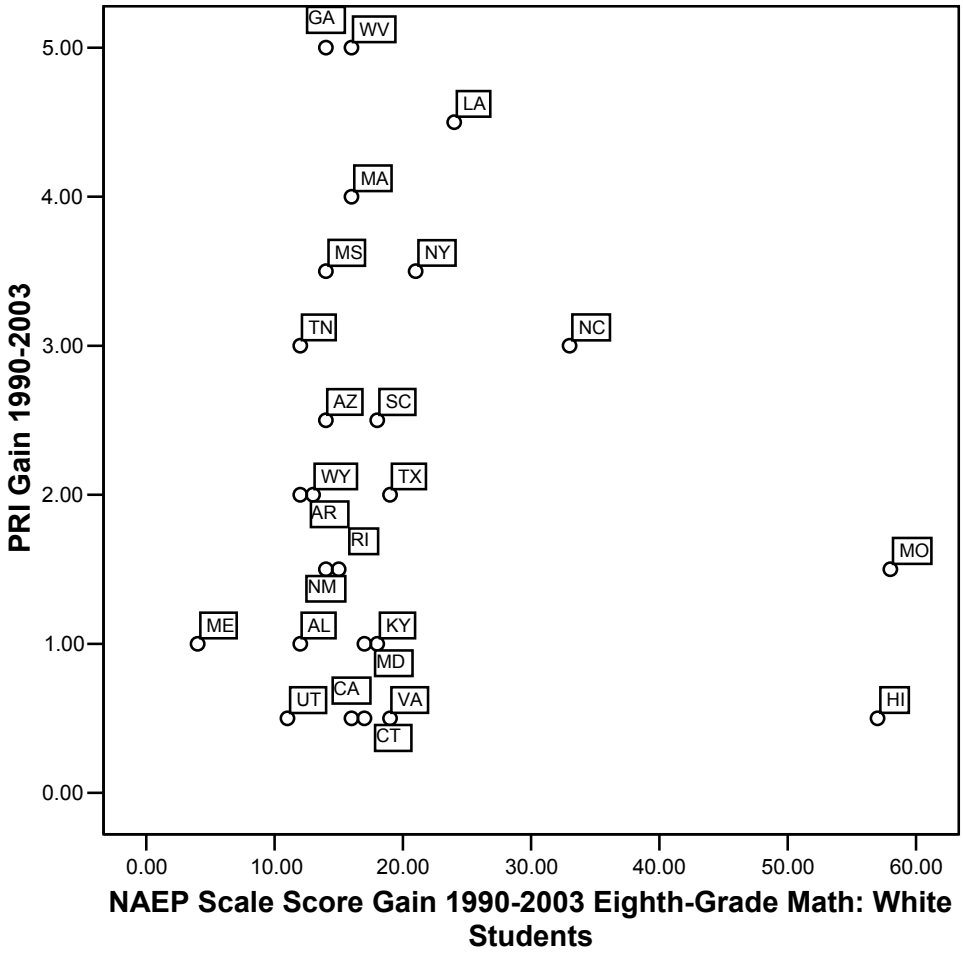
Across all students, there is a positive but moderate relationship between eighth-grade math gains and threat rating change (1990-2003). A scatter plot of the overall relationship of NAEP gain and threat change reveals a potential outlier (again, North Carolina) (see Figure 8). A follow-up correlation eliminating North Carolina from the equation changes the relationship only slightly (from $r = .283$ to $r = .268$).

Figure 8: Scatter Plot of Eighth-Grade Math NAEP Gain and Threat Rating Change: All Students



When disaggregated by student achievement, it can be seen that the relationship among achievement and pressure is virtually non existent for Hispanic and White students, but moderately strong for African American students. A scatter plot of the correlation among white students also reveals two outliers (see Figure 9). A correlation eliminating these two outliers (Hawaii and Missouri) change the relationship from r = -.106 to r = .269. There are no conspicuous outliers for the Hispanic or African American student data (see Appendix E).

Figure 9: Scatter Plot of Eighth-Grade Math NAEP Gain and Threat Rating Change: White Students



Reading

Correlations between NAEP gain and change in pressure for fourth- and eighth-grade reading are all relatively weak. A series of scatter plots also reveal no obvious outliers (see Appendix E).

Part III: Relationship of Change in PRI and Change in NAEP Achievement for "Cohorts" of Students

We wanted to see if changes in high-stakes testing pressure were related to changes in achievement among cohorts of students. Table 24 presents these results for math and Table 25 displays them for reading.

Table 24: Correlations of Cohort Achievement Gains and PRI Change: Math

	1992 fourth-grade NAEP*	1996 eighth-grade NAEP*	Cohort change (1992-1996)	1992-1996 PRI change	1996 fourth-grade NAEP*	2000 eighth-grade NAEP*	Cohort change (1996-2000)	1996-2000 PRI change
1992 fourth-grade NAEP*	1.000							
1996 eighth-grade NAEP*	0.960	1.000						
Cohort change (1992-1996)	0.230	0.493	1.000					
1992-1996 PRI change	-0.166	-0.111	**0.131**	1.000				
1996 fourth-grade NAEP*	0.893	0.918	0.419	0.095	1.000			
2000 eighth-grade NAEP*	0.877	0.918	0.466	0.057	0.916	1.000		
Cohort change (1996-2000)	0.308	0.355	0.278	-0.058	0.182	0.561	1.000	
1996-2000 PRI change	-0.197	-0.205	-0.102	-0.133	-0.116	-0.109	**-0.028**	1.000

* Denotes no accommodations.

Correlations between cohort achievement gains in math and threat changes are displayed in bold in Table 24 and reflect a weak but positive relationship for the 1992-1996 cohort, and an even weaker but negative relationship for the 1994-1998 cohort.

Table 25: Correlations of Cohort Achievement Gains and PRI Change: Reading

	1994 fourth-grade reading NAEP*	1998 eighth-grade reading NAEP*	Cohort NAEP gain 1994-1998	PRI change 1994-1998	1998 fourth-grade reading NAEP	2002 eighth-grade reading NAEP	Cohort NAEP gain 1998-2002	PRI change 1998-2002
1994 fourth-grade reading NAEP*	1.000							
1998 eighth-grade reading NAEP*	0.885	1.000						
Cohort NAEP gain 1994-1998	-0.676	-0.255	1.000					
PRI change 1994-1998	0.126	0.069	**-0.152**	1.000				
1998 fourth-grade reading NAEP	0.932	0.914	-0.489	0.059	1.000			
2002 eighth-grade reading NAEP	0.861	0.882	-0.391	0.107	0.869	1.000		
Cohort NAEP gain 1998-2002	-0.535	-0.465	0.374	0.046	-0.656	-0.196	1.000	
PRI change 1998-2002	-0.158	-0.026	0.286	-0.121	-0.109	-0.021	**0.184**	1.000

* Denotes no accommodations.

Correlations between cohort achievement gains in reading and threat changes are displayed in bold (Table 25) and reflect a weak but positive relationship for the 1998-2002 cohort, and a weak but negative relationship for the 1994-1998 cohort.

Part IV: Antecedent-Consequent Relationships Between Change in PRI and Change in NAEP Achievement

In our last set of analyses, we attempt to move closer to warranted conclusions about any causal relationship between high-stakes testing pressure and academic achievement. In these analyses we adopt a design that involves the correlation of changes in the PRI index with subsequent changes in NAEP scale score achievement changes. Since causes must precede their effects, the lack of any correlation of PRI change with NAEP change would significantly embarrass any claim of a causal link. Moreover, any form of regression analysis that ignores changes in putative causal variables and ignores time sequences of putative causes and effects is vulnerable to alternative explanations. For example, high PRI states may also be poor in ways not accounted for by the other variables entered into the regression equation. However, correlations with changes in the PRI index are far less confounded by unaccounted for "third variables." The combination of correlating the differences in measures of the putative causes and effect and staggering these differenced variables so that the cause is measured before the effect has a tradition in the literature of econometrics, where it is related to what is known as "Granger causality"—after Clive W. J. Granger[60] who was awarded the Nobel Prize in Economics in 2003—and has been applied with some success in the study of alcohol consumption and liver cirrhosis death[61] and the study of the economy and suicide deaths.[62]

First, we present a series of correlations between antecedent PRI change and subsequent NAEP scale score gains (non-cohort and across fourth- and eighth-grade math and reading overall and disaggregated by student ethnicity). Second, we examine the

same patterns but using cohort NAEP score gains. To illustrate our strategy, we focus on fourth-grade math. We began by identifying NAEP years of administration (for fourth-grade math they are 1992, 1996, 2000, and 2003). NAEP gains are then calculated for the following years: 1992-1996 (calculated as the difference of 1996 NAEP scale score and 1992 NAEP scale score), 1996-2000 and 2000-2003. Once these gain years were identified, we calculated corresponding antecedent PRI changes. For example, for NAEP gains of 1992-1996, PRI change was calculated across the previous four years of 1988-1992. Similarly, for NAEP gains of 1996-2000, we calculated the corresponding PRI change for the previous four years of 1992-1996. Lastly, for the NAEP gain of 2000-2003 we calculated the corresponding antecedent PRI change of the previous three years of 1997-2000.[63]

Cross-Sectional Causal Analyses

Our first set of causal analyses for fourth-grade math is presented in Table 26. All correlations between NAEP gain and previous pressure change are virtually nonexistent (see Table 27). Furthermore, a series of partial correlations holding exclusion rates constant do not change the nature of this outcome. Thus, for fourth-grade math achievement, the relationship between previous pressure increase and later NAEP achievement change is nonexistent.

Table 26: Correlations of PRI Change and NAEP Gains Across 1992-2003: Fourth-Grade Math—Non-Cohort

	PRI Change 1988-1992	NAEP Gain 1992-1996	% Excluded 1996****	PRI Change 1992-1996	NAEP Gain 1996-2000	% Excluded 2000	PRI Change 1996-2000	NAEP Gain 2000-2003	% Excluded 2003
PRI Change 1988-1992	1.000								
NAEP Gain 1992-1996	**-0.066***	1.000							
% Excluded 1996****	0.098	0.047	1.000						
PRI Change 1992-1996	-0.328	0.565	-0.048	1.000					
NAEP Gain 1996-2000	0.247	0.038	-0.019	**0.159****	1.000				
% Excluded 2000	0.053	0.319	0.602	0.160	0.149	1.000			
PRI Change 1996-2000	0.325	0.190	0.098	-0.151	0.112	0.137	1.000		
NAEP Gain 2000-2003	0.063	-0.297	0.081	0.060	0.124	-0.274	**0.142*****	1.000	
% Excluded 2003	0.212	0.336	0.237	0.150	0.092	0.463	0.246	0.028	1.000

Partial correlation results: * -.072; ** .138; *** .140
**** Denotes no accommodations.

The same set of analyses disaggregating the data by student ethnicity is calculated (see Table 27). As can be seen, earlier pressure changes are not related to achievement changes for African American, Hispanic or White students earlier in the 1990s. However, as the decade progresses, the relationship between antecedent pressure increases and later achievement gains strengthens. Specifically, for all subgroups, pressure change in the later half of the 1990s is strongly associated with most recent 2000-2003 NAEP gains.

Table 27: Correlations of PRI Change and NAEP Gains Across 1992-2003 by Student Ethnicity: Fourth-Grade Math—Non-Cohort

	A	B	C	D	E	F	G	H	I	J	K	L
A: PRI Change 1988-1992	1.000											
B: NAEP Gain 1992-1996* African American	**-0.087**	1.000										
C: NAEP Gain 1992-1996* Hispanic	**0.250**	0.226	1.000									
D: NAEP Gain 1992-1996* White	**0.000**	0.361	0.412	1.000								
E: PRI Change 1992-1996	0.042	0.440	0.539	0.906	1.000							
F: NAEP Gain 1996-2000* African American	-0.151	0.350	0.048	0.247	**0.427**	1.000						
G: NAEP Gain 1996-2000* Hispanic	-0.271	0.379	-0.427	0.131	**0.182**	0.368	1.000					
H: NAEP Gain 1996-2000* White	-0.090	0.319	-0.134	-0.362	**-0.159**	0.436	0.306	1.000				
I: PRI Change 1996-2000	-0.031	-0.180	-0.233	-0.409	-0.436	-0.268	0.258	0.130	1.000			
J: NAEP Gain 2000-2003 African American	-0.013	0.242	0.293	0.107	0.187	0.001	0.224	0.316	**0.374**	1.000		
K: NAEP Gain 2000-2003 Hispanic	0.084	0.076	0.053	-0.138	-0.135	-0.243	-0.098	0.306	**0.418**	0.423	1.000	
L: NAEP Gain 2000-2003 White	-0.090	0.005	-0.153	-0.360	-0.228	-0.049	0.362	0.287	**0.730**	0.278	0.216	1.000

* Denotes no accommodations.

In our next set of analyses, the relationship between pressure change and NAEP gain for eighth-grade math achievement is examined. As can be seen in Table 28, there is a moderate and positive relationship between earlier pressure change and later NAEP gain for the years 1990-1992 and 1996-2000 and a weak, but positive one for the years of 2000-2003. By contrast, there is a moderate but *negative* relationship between pressure change and NAEP gain for the 1992-1996 year. Corresponding partial correlations do not change this outcome significantly.

Table 28: Correlations of PRI Change and NAEP Gains 1990-2003 All Students: Eighth-Grade Math—Non-Cohort

	A	B	C	D	E	F	G	H	I	J	K	L
A: PRI Change 1988-1990	1.000											
B: NAEP Gain 1990-1992***	**0.223***	1.000										
C: % Excluded 1992	0.199	0.027	1.000									
D: PRI Change 1988-1990	0.464	-0.010	-0.058	1.000								
E: NAEP Gain 1992-1996***	-0.209	0.248	-0.008	**-0.297****	1.000							
F: % Excluded 1996	0.245	0.111	0.699	0.087	-0.010	1.000						
G: PRI Change 1992-1996	-0.201	-0.004	-0.034	-0.328	0.621	-0.033	1.000					
H: NAEP Gain 1996-2000***	0.193	0.381	-0.066	0.241	0.359	-0.077	**0.411*****	1.000				
I: % Excluded 2000	0.073	0.534	0.253	-0.240	0.449	0.418	0.317	0.378	1.000			
J: PRI Change 1997-2000	0.426	0.011	0.437	0.384	-0.099	0.217	-0.085	0.247	0.101	1.000		
K: NAEP Gain 2000-2003**	0.132	-0.332	0.103	0.267	-0.582	-0.180	-0.118	-0.114	-0.453	**0.195******	1.000	
L: % Excluded 2003	-0.008	-0.382	-0.507	0.006	0.107	-0.325	0.256	0.020	-0.129	-0.365	0.157	1.000

Partial correlations: * = .222; ** = .299; *** = .331; **** = .26

***** Denotes no accommodations.

A series of correlations between pressure change and eighth-grade math gains by student ethnicity are presented in Table 29. Across all years, there is no relationship between pressure and African American student NAEP score gains. Among Hispanic students, pressure has no bearing on subsequent achievement in the early 1990s (1990-1992) or in the most recent round of NAEP testing (2000-2003). By contrast, there is a moderate but positive relationship between pressure and NAEP gains for the years 1992-1996 and 1996-2000. Among White students, pressure and NAEP change are inconsistently related. For the years 1992-1996 and 2000-2003 there is a weak but negative relationship between pressure and NAEP gains. By contrast for remaining years of 1990-1992 and 1996-2000 there is a moderate but positive relationship.

Table 29: Correlations of PRI Change and NAEP Gains by Student Ethnicity: Eighth-Grade Math—Non-Cohort

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A: PRI Change 1988-1990	1.000															
B: NAEP Gain 1990-1992* African American	**0.161**	1.000														
C: NAEP Gain 1990-1992* Hispanic	**0.161**	0.317	1.000													
D: NAEP Gain 1990-1992* White	**0.300**	0.462	0.550	1.000												
E: PRI Change 1988-1992	0.464	0.105	0.087	0.058	1.000											
F: NAEP Gain 1992-1996* African American	-0.056	-0.011	0.140	0.033	**-0.114**	1.000										
G: NAEP Gain 1992-1996* Hispanic	-0.275	0.313	0.158	0.154	**0.245**	0.090	1.000									
H: NAEP Gain 1992-1996* White	-0.271	0.183	0.347	0.066	**-0.176**	0.086	0.426	1.000								
I: PRI Change 1992-1996	-0.201	0.121	0.259	-0.085	-0.328	0.305	0.158	0.489	1.000							
J: NAEP Gain 1996-2000* African American	0.130	-0.040	0.241	0.330	0.102	-0.655	-0.012	0.136	**-0.021**	1.000						
K: NAEP Gain 1996-2000* Hispanic	0.285	0.211	0.303	0.282	-0.049	0.387	0.041	0.344	**0.314**	0.166	1.000					
L: NAEP Gain 1996-2000* White	0.118	0.404	0.561	0.440	0.231	0.073	0.367	0.193	**0.334**	0.354	0.421	1.000				
M: PRI Change 1997-2000	0.426	0.011	-0.053	0.101	0.384	0.154	0.011	-0.266	-0.085	0.155	0.272	0.221	1.000			
N: NAEP Gain 2000-2003 African American	0.212	0.286	0.186	0.128	0.340	0.336	0.331	-0.065	0.152	-0.335	0.157	0.403	**-0.004**	1.000		
O: NAEP Gain 2000-2003 Hispanic	-0.146	-0.260	-0.292	-0.409	0.000	0.106	-0.232	-0.249	-0.186	-0.303	-0.313	-0.293	**-0.085**	0.025	1.000	

P: NAEP Gain 2000-2003 White	-0.159	-0.234	-0.150	-0.376	-0.198	0.095	-0.106	-0.111	0.024	-0.167	-0.224	-0.157	**-0.154**	-0.139	0.685	1.000

* Denotes no accommodations.

In our next set of analyses, pressure and NAEP gains for fourth-grade reading achievement are analyzed (Table 30). Again, data suggest an inconsistent effect of pressure on later achievement. For example, there is a moderate relationship between pressure change and NAEP gain in the early to mid-1990s, but in one case it is a negative relationship (1992-1994) and in the other it is positive (1994-1998). In later years, there is no relationship between changes in high-stakes testing pressure and subsequent NAEP achievement gains.

Table 30: Correlations of PRI Change and NAEP Gains 1992-2003 Fourth-Grade Reading

	A	B	C	D	E	F	G	H	I	J	K	L
A: PRI Change 1990-1992	1.000											
B: NAEP Gain 1992-1994***	**-0.313***	1.000										
C: % Excluded 1994	-0.194	-0.180	1.000									
D: PRI Change 1990-1994	0.879	-0.161	-0.130	1.000								
E: NAEP Gain 1994-1998*	0.159	-0.322	0.276	**0.143****	1.000							
F: % Excluded 1998	-0.002	-0.320	0.668	-0.065	0.713	1.000						
G: PRI Change 1994-1998	-0.192	0.142	-0.149	-0.292	-0.128	-0.043	1.000					
H: NAEP Gain 1998-2002	0.198	-0.453	-0.026	0.184	0.035	-0.021	**0.021*****	1.000				
I: % Excluded 2002	-0.086	-0.265	0.178	-0.080	0.235	0.468	0.096	0.359	1.000			
J: PRI Change 1994-1998	-0.192	0.142	-0.149	-0.292	-0.128	-0.043	1.000	0.021	0.096	1.000		
K: NAEP Gain 1998-2003	0.123	-0.405	-0.124	0.144	0.085	-0.073	0.125	0.786	0.266	**0.125******	1.000	
L: % Excluded 2003	-0.223	-0.269	0.352	-0.143	0.330	0.431	-0.001	0.114	0.772	-0.001	0.163	1.000

Partial correlations: * = .361; ** = .270; *** =.014; **** = .127

***** Denotes no accommodations.

We followed up these analyses looking at fourth-grade reading trends with earlier pressure and disaggregated by student ethnicity (see Table 31). Our results reveal no consistent pattern in the effect of pressure on achievement. Among African American students, NAEP gains in the early 1990s and antecedent pressure change are strongly but negatively associated (1992-1994). But over time, the relationship between pressure and achievement is virtually nonexistent. Similarly, there is no consistent pattern of relationships between pressure and achievement change among Hispanic or White students. In fact, most of the relationships are virtually nonexistent with the exception of pressure change and 1998-2002 NAEP gain among Hispanic students (which is negative, -.303) and 1998-2003 NAEP gain among White students (.280).

Table 31: Correlations of PRI Change and NAEP Gains by Student Ethnicity: Fourth-Grade Reading—Non-Cohort

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A: PRI Change 1990-1992	1.000															
B: NAEP Gain 1992-1994* African American	**-0.378**	1.000														
C: NAEP Gain 1992-1994* Hispanic	**0.118**	0.155	1.000													
D: NAEP Gain 1992-1994* White	**-0.047**	0.051	0.158	1.000												
E: PRI Change 1990-1994	0.879	-0.348	0.059	0.013	1.000											
F: NAEP Gain 1994-1998* African American	0.139	-0.571	-0.362	0.201	**0.132**	1.000										
G: NAEP Gain 1994-1998* Hispanic	-0.039	-0.245	-0.576	0.242	**0.059**	0.468	1.000									
H: NAEP Gain 1994-1998* White	0.141	-0.335	-0.020	-0.064	**0.044**	0.255	0.189	1.000								
I: PRI Change 1994-1998	-0.192	0.104	0.010	0.086	-0.292	-0.164	-0.178	-0.170	1.000							
J: NAEP Gain 1998-2002 African American	0.242	-0.156	0.235	-0.023	0.267	-0.121	0.044	0.354	**0.082**	1.000						
K: NAEP Gain 1998-2002 Hispanic	-0.272	0.147	-0.044	-0.046	-0.297	-0.014	-0.109	0.053	**-0.303**	0.175	1.000					
L: NAEP Gain 1998-2002 White	0.008	-0.228	-0.070	-0.212	0.083	0.094	-0.024	-0.117	**0.119**	0.599	0.173	1.000				
M: PRI Change 1994-1998	-0.192	0.104	0.010	0.086	-0.292	-0.164	-0.178	-0.170	1.000	0.082	-0.303	0.119	1.000			
N: NAEP Gain 1998-2003 African American	0.198	-0.303	0.318	-0.012	0.320	-0.043	-0.134	0.299	0.013	0.760	0.030	0.464	**0.013**	1.000		
O: NAEP Gain 1998-2003 Hispanic	-0.261	0.165	-0.028	-0.027	-0.296	-0.091	-0.110	-0.029	-0.113	0.133	0.810	0.205	**-0.113**	-0.008	1.000	

P: NAEP Gain 1998-2003 White	0.128	-0.174	0.054	-0.230	0.174	0.137	-0.129	-0.140	0.280	0.353	-0.252	0.742	**0.280**	0.406	-0.060	1.000

* Denotes no accommodations.

Lastly, patterns in antecedent pressure changes and subsequent NAEP change for eighth-grade reading achievement are examined (see Table 32). There is no evidence of a relationship between pressure and achievement for eighth-grade reading on average or when data are disaggregated by student ethnicity (see Table 33).

Table 32: Correlations of PRI Change and NAEP Gains Across 1992-2003 Eighth-Grade Reading—Non-Cohort

	PRI Change 1994-1998	NAEP Gain 1998-2002	% Excluded 2002	PRI Change 1993-1998	NAEP Gain 1998-2003	% Excluded 2003
PRI Change 1994-1998	1.000					
NAEP Gain 1998-2002	**0.085***	1.000				
% Excluded 2002	-0.013	0.292	1.000			
PRI Change 1993-1998	0.849	0.202	-0.020	1.000		
NAEP Gain 1998-2003	0.008	0.838	0.002	**0.102****	1.000	
% Excluded 2003	0.161	0.220	0.821	0.168	-0.066	1.000

Partial correlation: * = .093; ** = .115

Table 33: Correlations of PRI Change and NAEP Gains by Student Ethnicity: Eighth-Grade Reading—Non-Cohort

	PRI Change 1994-1998	NAEP Gain 1998-2002 African American	NAEP Gain 1998-2002 Hispanic	NAEP Gain 1998-2002 White	% Excluded 2002	PRI Change 1993-1998	NAEP Gain 1998-2003 African American	NAEP Gain 1998-2003 Hispanic	NAEP Gain 1998-2003 White	% Excluded 2003
PRI Change 1994-1998	1.000									
NAEP Gain 1998-2002 African American	**0.038**	1.000								
NAEP Gain 1998-2002 Hispanic	**0.149**	0.302	1.000							
NAEP Gain 1998-2002 White	**0.077**	0.317	0.146	1.000						
% Excluded 2002	-0.013	-0.030	0.220	0.384	1.000					
PRI Change 1993-1998	0.849	0.092	0.167	0.241	-0.020	1.000				
NAEP Gain 1998-2003 African American	-0.168	0.367	0.036	0.364	-0.047	**0.003**	1.000			
NAEP Gain 1998-2003 Hispanic	0.176	0.193	0.935	0.180	0.131	**0.157**	0.000	1.000		
NAEP Gain 1998-2003 White	0.109	0.346	0.194	0.701	0.342	**0.123**	0.100	0.280	1.000	
% Excluded 2003	0.161	-0.024	0.170	0.234	0.821	0.168	-0.120	0.011	0.208	1.000

Cohort Causal Analyses

In this last section, we present a series of correlations between antecedent changes in pressure associated with high-stakes testing and subsequent NAEP gains by student cohorts (i.e., "cohort" analyses follow the achievement trends of students as they progress

from fourth to eighth grade[64]). For these analyses, cohort NAEP gains are calculated as: [eighth-grade achievement year *i*] – [fourth-grade achievement year (*i* - 4)]. Table 34 presents the results for math. As can be seen there is a strong and negative relationship between 1988-1992 PRI change and 1992-1996 cohort achievement gain. Subsequently, there is no relationship between pressure (1992-1996) and cohort change (1996-2000).

Table 34: Correlations of PRI Change and Cohort Math NAEP Gains

	PRI Change 1988-1992	NAEP Cohort Change 1992-1996	% Excluded 1996 Eighth-Grade Math	PRI Change 1992-1996	NAEP Cohort Change 1996-2000	% Excluded 2000 Eighth-Grade Math
PRI Change 1988-1992	1.000					
NAEP Cohort Change 1992-1996	**-0.369***	1.000				
% Excluded 1996 Eighth-Grade Math	0.087	0.131	1.000			
PRI Change 1992-1996	-0.328	0.131	-0.033	1.000		
NAEP Cohort Change 1996-2000	0.046	0.278	-0.087	**-0.058****	1.000	
% Excluded 2000 Eighth-Grade Math	-0.240	0.446	0.418	0.317	0.356	1.000

Partial correlations: * = -.385; ** = -.193

Follow-up analyses of these relationships and by student ethnicity are presented in Table 35. Results suggest there is no relationship between pressure and math achievement for Hispanic and White students and only a very moderate one for African American students.

Table 35: Correlations of PRI Change and Cohort Math NAEP Gains by Student Ethnicity

	A	B	C	D	E	F	G	H	I	J
A: PRI Change 1988-1992	1.000									
B: NAEP Cohort Change 1992-1996 African American	**0.214**	1.000								
C: NAEP Cohort Change 1992-1996 Hispanic	**0.193**	0.434	1.000							
D: NAEP Cohort Change 1992-1996 White	**0.130**	0.166	0.211	1.000						
E: % Excluded 1996 Eighth-Grade Math	0.087	0.510	0.293	0.163	1.000					
F: PRI Change 1992-1996	-0.328	0.297	0.018	-0.035	-0.033	1.000				
G: NAEP Cohort Change 1996-2000 African American	0.256	0.918	0.296	0.218	0.417	**0.213**	1.000			
H: NAEP Cohort Change 1996-2000 Hispanic	0.190	0.574	0.827	0.383	0.320	**0.126**	0.434	1.000		
I: NAEP Cohort Change 1996-2000 White	0.235	0.242	0.249	0.947	0.124	**-0.065**	0.307	0.445	1.000	
J: % Excluded 2000 Eighth-Grade Math	-0.240	0.312	0.156	0.187	0.418	0.317	0.332	0.303	0.185	1.000

PRI change and cohort trends for reading achievement are presented in Table 36. There is no relationship between pressure and reading gains. Similarly, follow-up analyses by student ethnicity (Table 37) reveal no consistent pattern of effect of pressure on achievement.

Table 36: Correlations of PRI Change and Cohort Reading NAEP Gains

	PRI Change 1990-1994	Cohort NAEP Change 1994-1998	% Excluded 1998	PRI Change 1994-1998	NAEP Cohort Change 1998-2002	% Excluded 2002
PRI Change 1990-1994	1.000					
Cohort NAEP Change 1994-1998	**0.104***	1.000				
% Excluded 1998	0.047	0.667	1.000			
PRI Change 1994-1998	-0.292	-0.152	-0.002	1.000		
NAEP Cohort Change 1998-2002	0.355	0.374	0.248	**0.046****	1.000	
% Excluded 2002	0.081	0.387	0.621	0.076	0.490	1.000

Partial correlations: * = .098; ** = .010

Table 37: Correlations of PRI Change and Cohort Reading NAEP Gains by Student Ethnicity

	PRI Change 1990-1994	Cohort NAEP Change 1994-1998 African American	Cohort NAEP Change 1994-1998 Hispanic	Cohort NAEP Change 1994-1998 White	PRI Change 1994-1998	NAEP Cohort Change 1998-2002 African American	NAEP Cohort Change 1998-2002 Hispanic	NAEP Cohort Change 1998-2002 White
PRI Change 1990-1994	1.000							
Cohort NAEP Change 1994-1998 African American	**0.269**	1.000						
Cohort NAEP Change 1994-1998 Hispanic	**-0.295**	-0.017	1.000					
Cohort NAEP Change 1994-1998 White	**-0.099**	0.212	0.145	1.000				
PRI Change 1994-1998	-0.292	0.150	-0.184	-0.143	1.000			
NAEP Cohort Change 1998-2002 African American	0.297	0.859	-0.212	0.279	**0.092**	1.000		
NAEP Cohort Change 1998-2002 Hispanic	-0.357	-0.158	0.814	0.191	**-0.242**	-0.141	1.000	
NAEP Cohort Change 1998-2002 White	0.286	0.170	-0.410	0.234	**0.113**	0.366	-0.246	1.000

Discussion

Replication of Carnoy and Loeb

Some of our findings replicate those reported by Carnoy and Loeb. For example, when our rating system was substitute for theirs, there was a strong association between state composition and population, and pressure associated with accountability. It seems relatively clear that larger states and those with a greater proportion of minority students tend to implement accountability systems that exert a greater level of pressure. But, when Carnoy and Loeb examined the relationship of students' National Assessment of Education Progress (NAEP) test performance from the early 1990s with the strength of accountability implementation later, their only significant finding was the negative association between fourth-grade White students' math performance and later accountability implementation. By contrast, our analysis revealed a *positive* relationship between earlier African American student math achievement and pressure but a *negative* one between the change in the percent at or above basic in fourth-grade reading (1992-1994) and pressure.

In their second regression model, Carnoy and Loeb found that math gains were significantly associated with accountability strength—especially among eighth graders. Using our pressure rating index (PRI), there was a positive relationship between eighth-grade NAEP gains and PRI; however, the strength of that relationship depended on the NAEP indicator and whether exclusion rates were partialed out of the correlation. When the change in the percent of students achieving at or above basic and among all students (1996-2000) was the indicator, the correlation with PRI was significant and positive at

.446. However, a partial correlation holding NAEP 2000 exclusion rates constant reduced this relationship to essentially zero: .026. By contrast, when NAEP scale scores were used, the relationship between achievement gains (again among all students, 1996-2000) and our index of pressure was also positive, but slightly weaker at .372 (with a partial correlation of .351). When disaggregated by ethnicity, the change in the percent of students at or above basic (1996-2000) and PRI is significant (.456) for Black students, but non-existent for White or Hispanic students. Thus, among eighth graders, and especially among African American eighth-graders, pressure seems to be positively related to increases in achievement. Among fourth graders, there was a positive relationship between change in percent at or above basic (1996-2000) math achievement and PRI among all students and when the data are disaggregated by ethnicity. But, the strength of those relationships was lower than what was found for eighth grade (ranging from .184 - .327).

These findings replicate what Carnoy and Loeb and others have found[65]—that pressure is related to increases in math NAEP performance later in the 1990s. This finding emerges more strongly for eighth-grade math performance than it does for fourth-grade performance, and for African American students than any other ethnic subgroup. It is hard to draw any meaningful conclusions from these findings because they are correlational in nature. Further, there is evidence that students are excluded at higher rates during post testing which raises questions as to the validity of academic "gain" scores.

Progression

It was surprising to find a positive correlation between our index of pressure and eighth-ninth-grade progression. We would have predicted, as Carnoy and Loeb found, that pressure and eighth-ninth grade progression were unrelated. Still, it was not surprising that consistent with what others have found,[66] pressure is negatively associated with the likelihood that students will progress into 12th grade. Thus, it may be that increasing pressure leads to greater numbers of students dropping out or being held back later in school. However, this conclusion is drawn with caution because, as others have noted, [67] the use of enrollment figures as a proxy for grade progression does not account for enrollment changes due to migration or movement from school to school.

PRI Change and NAEP Gains

In our second set of analyses, a series of correlations were calculated to examine the pattern of relationships among NAEP gains and pressure change, both over the same time period and based on an antecedent-consequent design. Our correlations of NAEP gains and PRI change across the same time period (1990-2003) across fourth- and eighth-grade levels and for both math and reading in aggregate and disaggregated by student ethnicity (Table 22) revealed mostly positive but weak correlations (the largest positive correlation was .383). But all correlations (among aggregated achievement scores) decreased when NAEP exclusion rates were held constant. This set of analyses suggests that between the first administration of NAEP (state level) and the most recent, the corresponding change in pressure was only slightly related to math achievement gains and only for certain subgroups (e.g., fourth-grade Hispanic and eighth-grade African American student achievement). But, by dramatic contrast pressure increases were

unrelated to reading gains at the fourth- or eighth-grade levels and among all ethnic student subgroups.

Our strongest findings rest in the antecedent-consequent analyses. The data summarized in Table 38 represent *averaged* instances of correlating antecedent PRI changes with subsequent NAEP scale score changes for both cohort and non-cohort analyses. These averaged correlations suggest that previous increases in pressure do not cause later increases in achievement. However, a review of the underlying constituent correlations represented in this table unmasks a subtle, but important, pattern. In Table 39, all possible antecedent-consequent correlations we found across each student ethnic subgroup are listed (from Tables 27, 29, 31, and 33). Of particular note in Table 39 is the fact that for the four largest (in absolute value) correlations (see bolded entries at the bottom of the table) obtained in all the antecedent-consequent analyses, all four are for *fourth-grade math, non-cohort analyses*. Moreover, three of these four correlations (.73, .42, .37) emanated from PRI changes that occurred during the last half of the 1990s. If the four largest correlation coefficients are removed, the remaining 35 coefficients average 0.05 and are fairly evenly distributed around zero with a standard deviation of 0.17, which is not far off of the standard error of correlations based on an *n* of 25 when the population value is zero.

Table 38: Antecedent—Consequent Relationships (Corr. Coeff) Between PRI Changes and NAEP Gains by Subject, Grade, Ethnicity, & Design (Non-Cohort vs. Cohort)

	Non-Cohort Analysis		Cohort Analysis (Grades 4-8)	
	Reading	**Math**	**Reading**	**Math**
African American Grade 4	.04	.24	.18	.21
African American Grade 8	.02	.00		
Hispanic Grade 4	-.06	.30	-.27	.16
Hispanic Grade 8	.15	.16		
White Grade 4	.10	.19	.07	.03
White Grade 8	.10	.08		

Table 39: Antecedent—Consequent Change Correlations for Various Subjects, Ethnicities, and Grades: Non-Cohort Analyses

-0.38
-0.30
-0.18
-0.16
-0.15
-0.11
-0.11
-0.09
-0.09
-0.05
-0.02
0.00
0.00
0.00
0.01
0.04
0.04
0.06
0.08
0.08
0.12
0.12
0.12
0.13
0.15
0.16
0.16
0.16
0.18
0.25
0.25
0.28
0.30
0.31
0.33
0.37
0.42
0.43
0.73

The pattern of these correlations speaks to the validity of the conclusion that we have indeed uncovered a causal link between high-stakes testing pressure and student achievement but only with respect to fourth-grade math, non-cohort trends. The four strong correlations noted in Table 39 appear under the following circumstances: Fourth-grade math, non-cohort analysis. It is significant that the strongest relationships were observed under these circumstances and not others (e.g., fourth- or eighth-grade reading, or even cohort analyses for fourth- or eighth-grade math). The difference between a NAEP gain score for a cohort analysis vs. a non-cohort analysis is that in the former case, the achievement of students is tracked from grade 4 to grade 8 *across intermediate grades math curricula*. In the latter case—non-cohort analysis—the achievement of one year's grade 4 students is compared to a subsequent year's grade 4 achievement *on grade 4 math curriculum (or more likely, grades 1-4 math curricula)*. The math curriculum in the primary grades (1-4) is more standardized across the nation than the math curriculum in the intermediate and middle school grades (5-8). Consequently, math achievement at these levels is more likely to be affected by drill and practice or teaching to a test because of the more "predictable" content.

These findings, in combination with our replication analyses and what others have found,[68] suggest that there is something about high-stakes testing that is related to math achievement—especially among fourth graders and particularly as accountability policies were enacted and enforced in the latter part of the 1990s and early 2000s. But, it is just as notable that high-stakes pressure has no relation to reading achievement at either the fourth- or eighth-grade levels or for any student subgroup. In the end, our findings (and lack of findings) lead us to the conclusion that high-stakes testing pressure might produce

effects only at the simplest level of the school curriculum: primary school arithmetic where achievement is most susceptible to being increased by drill and practice and teaching to the test.

Limitations and Future Directions

We recognize that our measurement of pressure, while innovative and comprehensive and an improvement over attempts made in previous research, is not without its limitations. In spite of all our efforts to create portfolios that describe as comprehensively as possible all state-level assessment and accountability activities, we are aware that in creating these, we had a potential bias toward the more negative aspects of "accountability." This inclination potentially influenced the structure, content, and flow of the portfolio documentation. Still, our approach was relatively robust for describing the rich variation of high-stakes testing implementation and impact.

The use of newspaper documentation for describing cultural events raises many questions of potential selectivity bias. In spite of our best efforts to minimize bias through a systematic news search and sampling process, the potential of news stories to assume a negative slant and to exaggerate stories they cover must be acknowledged. Still, by systematizing the sampling procedures for identifying stories to include in all portfolios, we hoped to eliminate, or at least dramatically reduce, between state differences in newspaper orientation (i.e., liberal versus conservative) and availability (Massachusetts had significantly more types of media covering educational accountability than a state such as Maine, for example). Further, recognizing that newspapers tend to favor negative accounts, we made a concerted effort to include any positive coverage that existed in the corpus.

Our procedures for identifying state-level pressure over time, and therefore the threat rating difference estimates (i.e., PRI change), should be augmented by the judgments of a greater number of experts. Although two judges conducted independent evaluations of a random selection of portfolios and compared their year-to-year pressure rating judgments, their rates of agreement across all years and all changes in pressure were moderate. Nonetheless, the primary consequence of unreliable ratings was not observed, i.e., some non-zero correlations of PRI changes with NAEP gains were observed which would not have been the case had the PRI ratings over time by the two judges been of very low reliability. In future studies, more work must be done to ensure agreement across all pressure ratings by state and year.

This study represents a significant contribution to the measurement of high-stakes testing pressure. Future studies could draw upon our characterizations to investigate the effects of pressure on other teacher/student outcomes. For example, is pressure associated with increases in students' antisocial behavior? Students (and teachers) under increased pressure might be induced to vent their anxiety and frustration in undesirable ways. This study represents a solid framework from which future students can examine the effects of pressure across a range of academic and social outcomes.

In light of the rapidly growing body of evidence of the deleterious unintended effects of high-stakes testing, and the fact that our study finds no convincing evidence that the pressure associated with high-stakes testing leads to increased achievement, there is no reason to continue the practice of high stakes testing. Thus, given (a) the unprofessional treatment of the educators who work in high-stakes testing situations, (b) the inevitable corruption of the indicators used in accountability systems where high-

stakes testing is featured, (c) data from this and other studies that seriously question whether the intended effects of high-stakes testing actually occur, and (d) the acknowledged impossibility of reaching the achievement goals set by the NCLB act in a reasonable time frame, there is every reason to ask for a moratorium on testing policies that force us to rely on high-stakes testing.

Notes & References

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

[2] American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

[3] Ryan, J. E. (2004). The perverse incentives of the No Child Left Behind Act. New York University Law Review, 79, 932-989.

[4] Jones, M. G., Jones, B. D., and Hargrove, T. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield, and Nichols, S. & Berliner, D. C. (2005). The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing. EPSL-0503-101-EPRU, March 2005.

[5] Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003, March). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved January 7, 2004, from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

[6] Noddings, N. (2001). Care and coercion in school reform. *Journal of Educational Change*, 2, 35-43

Noddings, N. (2002). High-stakes testing and the distortion of care. In J. L. Paul, C. D. Lavely, A. Cranston-Gringas, & E. L. Taylor (Eds.), *Rethinking professional issues in special education* (pp. 69-82). Westport, CT: Ablex Publishing Corporation.

[7] Linn, R. L. (2000). Assessments and accountability. *Education Researcher, 29* (2), 4–15

Messick, S. L. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Messick, S. L. (1995b). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

[8] Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press

Deci, E. L., Koestner, R., & Ryan, R. M. (2001, Spring). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research, 71*(1), 1-28

McCaslin, M., & Good, T. (1996). The informal curriculum. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 622-673). New York: Macmillan

Good, T. (1999) (Ed.) *Elementary School Journal (Special Issue), 99*(5); and Good, T. (2000), (Ed.) *Elementary School Journal (Special Issue), 100*(5).

[9] Jones, M. G., Jones, B. D., and Hargrove, T. (2003). *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield

Kohn, A. (2000). *The case against standardized: Raising the Scores, Ruining the Schools*. Portsmouth, NH: Heinemann; and Ryan, 2004.

[10] Amrein, A.L. & Berliner, D.C. (2002a). The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP Test results in states with high school graduation exams. Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved January 7, 2004, from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf

Amrein, A.L. & Berliner, D.C. (2002b). High-Stakes testing, uncertainty, and student learning Education Policy Analysis Archives, 10(18). Retrieved January 7, 2004 from http://epaa.asu.edu/epaa/v10n18/

[11] Braun, H. (2004, January). Reconsidering the impact of high-stakes testing. *Educational Policy Analysis Archives*, 12(1), 1-40. Retrieved February 5, 2005, from http://epaa.asu.edu/epaa/v12n1/

[12] Carnoy, M., & Loeb, S. (2002, Winter). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, 24 (4), 305-331.

[13] Braun, 2004, and Amrein & Berliner 2002a.

[14] Carnoy & Loeb, 2002.

[15] National Commission for Excellence in Education. (1983, April). *A nation at risk: The imperatives for educational reform*. Washington, DC: U.S. Department of Education, National Commission for Excellence in Education

[16] Berliner, D. C. & Biddle, B.J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley Publishing Company, Inc.

See also

Tyack, D. B., & Cuban, L. (1996). *Tinkering Toward Utopia: A Century of Public School Reform*. Cambridge, MA: Harvard University Press.

[17] The form and function of New York's Regents tests have changed over time. Previously, New York had a two-tiered diploma system where students received a "regular" diploma if they did not take/pass the Regents tests. By contrast, students who did pass the tests would receive a more prestigious diploma. More recently, however, students have to pass the Regents exam in order to receive any diploma.

[18] Amrein & Berliner, 2002a, Table 1.

[19] Sunderman, G. L., & Kim, J. (2004a, February). Inspiring vision, disappointing results: Four studies on implementing the No Child Left Behind Act. Cambridge, MA: The Civil Rights Project at Harvard University

Sunderman, G. L., & Kim, J. (2004b, February). *Expansion of federal power in American education: Federal-state relationships under No Child Left Behind Act, Year One*. Cambridge, MA: The Civil Rights Project at Harvard University.

[20] Haney, W. (2000, August). The Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved February 5, 2005 from http://epaa.asu.edu/epaa/v8n41/

See also

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000, October). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*(49). Retrieved February 22, 2005 from http://epaa.asu.edu/epaa/v8n49/

[21] See No Child Left Behind Act (NCLBA) of 2001 § 1001, 20 U.S.C. § 6301 (Statement of Purpose) Retrieved August 26, 2003, from http://www.ed.gov/legislation/ESEA02/107-110.pdf

See also

Center on Education Policy, From the capital to the classroom: State and Federal efforts to implement the No Child Left Behind Act, at iii (2003) (describing purpose of NCLBA), available at http://www.ctredpol.org/pubs/nclb_full_report_jan2003/nclb_full_report_jan2003.pdf

[22] e.g., Education Commission of the States (2002, March). *No state left behind: The challenges and opportunities of ESEA 2001*, Author. Retrieved February 6, 2005, from http://www.ecs.org/clearinghouse/32/37/3237.pdf.

See also

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003, July). *Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workbooks and U.S. Department of Education reviews under the No Child Left Behind Act of 2001.* Washington, DC: Council of Chief State School Officers.

[23] Linn, 2004; Sunderman, & Kim, 2004a, b; Noddings, 2001, 2002

Neill, M., Guisbond, L and Schaeffer, B., with Madison, J. & Legeros, L. (2004). *Failing Our Children, How "No Child Left Behind" Undermines Quality and Equity in Education and An Accountability Model that Supports School Improvement.* Cambridge, MA: Fairtest.

Orfield, G. & Kornhaber, M.L. (2001) (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* New York: The Century Foundation Press.

[24] For example, one *New York Times* article reported that in the fall of 2003, the 8,000 requests for school transfers were impossible to accommodate. See

Winerip, M. (2003, October 1). In 'No Child Left Behind,' a problem with math. *New York Times.*

Readers are referred to several studies outlining some of the problems associated with implementing No Child Left Behind. For example, see

Erpenbach, Forte-Fast, & Potts, 2003 and Education Commission of the States, 2002.

[25] Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003, March). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved January 7, 2004, from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

[26] Amrein & Berliner, 2002a, b; Braun, 2004.

[27] Amrein & Berliner, 2002a, p. 57.

[28] Rosenshine, B. (2003, August 4). High-Stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24) p. 4. Retrieved January 7, 2004 from http://epaa.asu.edu/epaa/v11n24/

[29] Amrein-Beardsley, A., & Berliner, D. (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 February 5, 2005, from http://epaa.asu.edu/epaa/v11n25/

[30] Exclusion rates are defined as those students excluded from the assessment because "school officials believed that either they could not participate meaningfully in the assessment or that they could not participate without assessment accommodations that the program did not, at the time, make available. These students fall into the general categories of students with disabilities (SD) and limited-English proficient students (LEP). Some identified fall within both of these categories." From

Pitoniak, M. J., & Mead, N. A. (2003, June). Statistical methods to account for excluded students in NAEP. Educational Testing Service, Princeton, NJ. Prepared for U.S. Department of Education; Institute of Education Sciences, and National Center for Education Statistics; p. 1. Retrieved February 14, 2005 from http://nces.ed.gov/nationsreportcard/pdf/main2002/statmeth.pdf

[31] Braun, 2004, p. 33.

[32] Braun, 2004, p. 33.

[33] Heubert, J.P. & Hauser, R.M., Eds. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press

Orfield, G., Losen, D., Wald, J., & Swanson, C. B. (2004). *Losing our future: How minority youth are being left behind by the graduation rate crisis.* Cambridge, MA: The Civil Rights Project at Harvard University. Contributors: Advocates for Children of New York, The Civil Society Institute.

See also

Reardon, S. F., & Galindo, C. (2002, April). *Do high-stakes tests affect students' decisions to drop out of school? Evidence from NELS.* Paper presented at the Annual Meetings of the American Educational Research Association, New Orleans

Clarke, M., Haney, W., & Madaus, G. (2000). *High Stakes Testing and High School Completion.* Boston, MA: Boston College, Lynch School of Education, National Board on Educational Testing and Public Policy.

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

[35] Braun, 2004; Rosenshine, 2003.

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

[37] Carnoy & Loeb, 2002, p. 311.

[38] Carnoy & Loeb, 2002, Appendix A.

[39] Carnoy & Loeb, 2002, p. 14.

[40] Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J., & Li, J. (2003, January). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states*. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved January 7, 2004, from http://www.bc.edu/research/nbetpp/statements/nbr1.pdf

Pedulla, et al., 2003.

[41] Amrein and Berliner, 2002a, Table 1.

[42] The Education Commission of States is a data warehouse initiated by James Conant over 30 years ago who believed that there should exist "a mechanism, by which each state knows exactly what the other states have done in each education area, and the arguments pro and con. We ought to have a way by which the states could rapidly exchange information and plans in all education matters from the kindergarten to the graduate schools of a university" (downloaded January 17, 2005, from http://www.ecs.org/ecsmain.asp?page=/html/aboutECS/WhatWeDo.htm).

The mission of ECS is to "help state leaders identify, develop and implement public policy for education that addresses current and future needs of a learning society. (Downloaded January 17, 2005 from, http://www.ecs.org/clearinghouse/28/32/2832.htm). More information on ECS and their database can be found online: http://www.ecs.org/

[43] ECS's database of state-level accountability laws and activities is probably the most accurate and comprehensive as of 2001.

[44] ECS, 2002; Amrein & Berliner, 2002a, Table 1.

[45] NAEP began disaggregating student achievement by state in 1990. Eighteen states participated in this assessment schedule since its inception and therefore have available a complete set of NAEP data on fourth- and eighth-grade students in math and reading. These are Alabama, Arizona, Arkansas, California, Connecticut, Georgia, Hawaii, Kentucky, Louisiana, Maryland, New Mexico, New York, North Carolina, Rhode Island, Texas, Virginia, West Virginia, and Wyoming. Seven states

are missing one assessment—eighth-grade math test from 1990. These are South Carolina, Massachusetts, Maine, Mississippi, Missouri, Tennessee, and Utah. All 25 states are the focus of this study.

[46] The first author inquired about how ECS obtained the information provided in their table. Personal correspondence revealed that the lead researcher in charge of maintaining this database on state-level accountability laws consulted a variety of sources including legal briefs, laws, discussions with state department of education representatives and state department of education websites.

[47] Altheide, D. L. (1996). *Qualitative media analysis*. Quantitative Research Methods, Volume 38. Thousand Oaks, CA: SAGE Publications, p. 2.

[48] Torgerson, W. S., (1960). *Theory and methods of scaling.* John Wiley and Sons, Inc: New York.

[49] Torgerson, 1960, pp. 159-160.

[50] As outlined in Torgerson, 1960, pp. 170-173.

[51] No systematic studies were conducted to test this assumption.

[52] We did not include these same figures for the year 1963 as Carnoy and Loeb did and therefore, did not conduct an exact replication of this regression model.

[53] Carnoy & Loeb, 2002.

[54] NAEP data downloaded from the National Center for Education Statistics website, http://nces.ed.gov/

[55] Both from the National Center for Education Statistics website, http://nces.ed.gov/

[56] Downloaded from US Census Bureau website, http://www.census.gov

[57] We requested from Carnoy the data set they used for their analysis to ensure exact replication. But, although they shared some information with us on their accountability rating index, we did not receive the exact data they used as predictor variables. Thus, our analysis does not represent an exact replication.

[58] Amrein & Berliner, 2002a

[59] All enrollment data downloaded from http://nces.ed.gov/

[60] See pp. 620 ff, Gujaratim D.N., (1995), *Basic Econometrics, 3rd Ed*. New York: McGraw Hill.

[61] Lynch, W., Glass, G.V., & Tran, Z.V. (1988). Diet, tobacco, alcohol and stress as causes of coronary heart disease: A longitudinal causal analysis. *Yale Journal of Biology and Medicine*. Vol. 61, 413-426.

[62] Webb, L.D., Glass, G.V, Metha, A. and Cobb, C. (2002). Economic Correlates of Suicide in the United States (1929-1992): A Time Series Analysis. *Archives of Suicide Research*, *6*(2), 93-101.

[63] We conducted the same set of analyses keeping the four-year intervals among PRI change constant. That is, we correlated NAEP gain of 2000-2003 with PRI change 1997-2000 and with PRI change 1996-2000. There were no important differences in any of these corresponding analyses. Therefore, for consistency we calculated PRI change in terms of the number of years NAEP change was calculated.

[64] A "cohort" is not a true cohort in the sense that we follow the same students from fourth to eighth grade. Rather, it is a proxy of a true cohort—following the achievement trends of students as they progress through the intermediary grades from fourth to eighth grade.

[65] Braun, 2004; Rosenshine, 2003

[66] Haney, W., Madaus, G., Abrams, L., Wheelock, A., Miao, J., & Gruia, I (2004, January). The education pipeline in the United States 1970-2000. National Board on Educational Testing and Public Policy. Chestnut Hill, MA: Boston College.

[67] Heubert & Hauser, 1999; and Haney et al., 2004.

[68] Braun, 2004 and Carnoy & Loeb, 2002.