Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Sherman Dorn
College of Education
University of South Florida

Volume 14 Number 30 **November 20, 2006** **ISSN 1068–2341**

Relationships between High-Stakes Testing Policies and Student Achievement after Controlling for Demographic Factors in Aggregated Data

Gregory J. Marchant, Sharon E. Paulson, and Adam Shunk
Ball State University

Citation: Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relationships between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Education Policy Analysis Archives, 14*(30). Retrieved [date] from http://epaa.asu.edu/epaa/v14n30/.

Abstract

With the mandate of *No Child Left Behind*, high-stakes achievement testing is firmly in place in every state. The few studies that have explored the effectiveness of high-stakes testing using NAEP scores have yielded mixed results. This study considered state demographic characteristics for each NAEP testing period in reading, writing, mathematics, and science from 1992 through 2002, in an effort to examine the relation of high-stakes testing policies to achievement and changes in achievement between testing periods. As expected, demographic characteristics and their changes were related significantly to most achievement outcomes, but high-stakes testing policies demonstrated few relationships with achievement. The few relationships between high-stakes testing and achievement or improvement in reading, writing, or science tended to appear only when demographic data were missing; and the minimal relationships with math achievement were consistent with findings in previous research. Considering the cost and potential unintended negative consequences, high-stakes testing policies seem to provide a questionable means of improving student learning.

Keywords: high-stakes tests; national competency tests; evaluation; educational policy; academic standards.

Relaciones entre las políticas educativas de pruebas de alto impacto (high-stakes testing) y el logro académico de los estudiantes, después de controlar factores demográficos en datos agregados

Resumen
Con el mandato "Ningún Niño(a) Rezagado" (No Child Left Behind), las pruebas de alto impacto sobre el logro académico de los estudiantes se han plantado firmemente en cada estado. Los pocos estudios que han explorado la efectividad de los exámenes de alto impacto usando los resultados de la Evaluación Nacional del Progreso Educativo (NAEP) han producido resultados mixtos. En un esfuerzo por examinar la relación de las políticas sobre pruebas de alto impacto con el logro académico y los cambios en el logro académico entre los períodos de pruebas, este estudio tomó en cuenta características demográficas de cada estado para cada período de pruebas de la NAEP en lectura, escritura, matemáticas y ciencias desde 1992 hasta 2002. Como se esperaba, las características demográficas y sus cambios, están significativamente relacionados con la mayoría de resultados de logro académico. Por otro lado, las políticas de pruebas de alto impacto demostraron pocas relaciones con el logro académico. Las escasas relaciones entre las pruebas de alto impacto y el logro o mejoramiento en lectura, matemáticas o ciencias aparecieron solamente cuando no se incluían datos demográficos; y las relaciones mínimas con los resultados en el área de matemáticas son consistentes con los resultados de investigaciones previas. Tomando en cuenta el costo y el potencial de producir consecuencias negativas, las políticas de exámenes de grandes consecuencias parecieran ofrecer un medio dudoso de mejorar los resultados de aprendizaje académico de los/as estudiantes.

Introduction

The American public education system is known for jumping on bandwagons of programs and curriculum innovations without a thorough evaluation of their effectiveness. The most public (at least in the educational research community) example of this problem emerged with evaluations of the Hunter Method. As whole states were incorporating these approaches into curriculum, training, and teacher evaluation, the costly program was evaluated and deemed ineffective (Slavin, 1989). One of the largest bandwagons that has been building speed for more than a decade, and received a giant push with the passing of the *No Child Left Behind* federal legislation, is the use of standardized achievement tests as accountability measures for states, districts, schools, teachers, and most importantly, students. When such accountability is associated with serious consequences, the tests are termed high-stakes (American Educational Research Association, 2000). Advocates believe that testing, and its associated standards and consequences, provide direction and motivation for achievement. However, this bandwagon has been called a Trojan horse (Corno, 2000; Paris, 2000) for the unintended negative consequences that have been identified (Amrein & Berliner, 2002c; Jones, Jones, & Hargrove, 2003; Marchant, 2005; Marchant & Paulson, 2005).

Currently all states have established testing programs to meet the *No Child Left Behind* (NCLB) mandate. However, it is still unclear whether these programs are indeed in the best interests of those who stand to gain or lose the most: America's public education students. Little research has been conducted to assess the influences of accountability systems on education in general and on student learning in particular. It is unclear at this point whether or not the intended goals of accountability systems are being realized (Amrein & Berliner, 2002a; Linn, 2000; Steinberg, 2003). Although NCLB is a federal mandate of accountability, its implementation varies greatly from state to state, with some states having had testing programs in place long before being required. The ultimate goal of accountability systems established by NCLB is to increase student achievement (such that every child will reach "proficiency" on state-determined achievement tests by academic year 2014–2015); therefore, one of the ways that the effectiveness of high-stakes testing policies could be assessed would be to compare students' achievement in states that had established accountability systems prior to the 2002 testing mandate to those that had not. The one database with comparable state-level achievement data for all 50 states is that of the National Assessment of Educational Progress (NAEP, also referred to as the Nation's Report Card; Jones & Olkin, 2004). Some studies have used the NAEP data to assess the effects of high-stakes testing on student achievement, with varying results (Amrein & Berliner, 2002a, 2002b; Braun, 2004; Carnoy & Loeb, 2002; Rosenshine, 2003; Nichols, Glass, & Berliner, 2006). However, differences exist in how each of these studies identified each state as high-stakes and what potentially confounding variables are included as controls.

In their study using NAEP data, Amrein and Berliner (2002a, 2002b) examined longitudinal changes in NAEP scores before and after states implemented high-stakes testing policies. Although they found that some states did show small gains, other states did not, and they concluded that there was no compelling evidence that the implementation of high-stakes testing improved student achievement. They argued that states could show gains in test scores by excluding certain groups of students (disabled or limited English proficient); however, they did not systematically control for specific student characteristics that may have biased the testing samples. In contrast, Rosenshine (2003) reported overall gains on the NAEP in states with high-stakes testing; however, he also did not control for any student demographic characteristics that may potentially be confounded with these gains. In their study of NAEP data, Carnoy and Loeb (2002) used recursive regression models that would correct for some of the limitations of previous studies. In particular, they created a measure of the strength of each state's accountability system (rather than just identifying the date of implementation of high-stakes testing) that included specific characteristics of individual states like funding, ethnicity, and population. They also controlled for the student characteristics most known to be confounded with achievement gains (ethnicity and inclusion/exclusion from testing). Their results showed positive relationships between states' level of accountability and gains in NAEP scores.

However, two major flaws still can be identified in this study. First, although they controlled for demographic characteristics of students, the authors failed to control for those characteristics known to be highly related to achievement outcomes: family income and parent education levels (see Eccles, 2005; Entwisle, Alexander, & Steffel Olsen, 2005; Sirin, 2005; and White, 1982, for discussions of hundreds of research studies confirming the relationships between family income or parent education and children's achievement). There is considerable variation among states on these family factors and it has been shown that even small changes in demographic characteristics of a testing sample can produce large changes in achievement (Marchant, 2005). Second, Carnoy and Loeb (2002) only assessed changes in math scores. The NAEP also measures reading, writing, and science. One cannot generalize findings regarding math as evidence that high-stakes testing policies lead to positive gains in all student achievement. In 2004, Braun reanalyzed relative gains for states on the NAEP mathematics tests while controlling for percentage of students excluded from the testings. His results favored high-stakes testing states when comparing the same grade levels (i.e., fourth grade at one testing period to fourth grade four years later), but his results favored low-stakes testing states when comparing a cohort of fourth graders to eighth graders four years later.

Most recently, Nichols and her colleagues (2006) replicated Carnoy and Loeb's (2002) work, but they added their own indicator of "pressure" of accountability and they assessed the results for reading achievement as well. Several of their conclusions are relevant to the current study. First, they concurred with Carnoy and Loeb that higher strength of accountability (particularly greater high stakes testing pressure) might reflect positively in math scores at the primary level (fourth grade), but they claimed that pattern existed only because primary math curricula may be more responsive than others to the drill and practice exhibited by schools under greater accountability pressure. Indeed, neither higher grade (eighth) math nor reading proficiency was affected by greater strength (or pressure) of accountability. The authors argued that even small gains in math among African American students may be difficult to interpret because exclusions from testing which increase at higher grade levels may call into question the validity of some findings. But this study also neglected to control for what Berliner (2005) considered the "600 pound gorilla"—family factors, namely family income; and some might argue that parent education is equally important (Marchant, 2005; Marchant & Paulson, 2005; Sirin, 2005). Any changes in achievement scores aggregated by state cannot be interpreted without first controlling for potential changes in the factors known to be most highly correlated with students' achievement outcomes.

The purpose of this study was to examine the relationship of states' high-stakes testing policies to students' NAEP achievement after controlling for differences in students' demographic characteristics known to be related to NAEP achievement: family income, parent education, ethnicity, and exclusion of disabled or limited English proficient students. Ethnicity and exclusion of disabled and LEP students were included because of researchers' recognition of their importance, reflected in other studies of high stakes testing, in particular those using NAEP data (e.g., Braun, 2004; Carnoy & Loeb, 2002; Nichols et al., 2006). Family income and parent education were added to the analyses to demonstrate the significant impact these factors also have in explaining achievement differences among states (Marchant, 2005). It was not the intention of the authors of this study to replicate either Carnoy and Loeb (2002) or Nichols et al. (2006). Instead, we used what we considered to be more conventional statistical techniques to demonstrate the importance of considering family income and parent education levels of test-takers in comparing groups (i.e., states) on aggregated achievement data. Indeed, we believed that we could affirm Nichols' and colleagues' contention that high stakes testing policies do little to produce higher achievement even when the stakes are highest.

The study assessed the influence of high-stakes testing policies using three indicators: whether or not high-stakes testing was implemented, the number of years high-stakes testing had been in place (both employed by Amrein and Berliner, 2002a, 2002b), and the strength of accountability index employed by Carnoy and Loeb (2002). These indicators were used for all four subject areas assessed by the NAEP: reading, writing, math, and science. Because individual student data on the NAEP were not available for this study, data were aggregated into "testing samples," defined as the group of students in each state, at each grade level (fourth and eighth grades), during each NAEP testing period (e.g., 1992, 1996, and 2000 for math; or 1994, 1998, and 2002 for reading). Although a total of 300 testing samples might be possible for each subject area (two grade levels x three testing periods x 50 states, including DC—South Dakota did not participate), some states did not administer the test at every testing period, some states did not administer the test to both grade levels, and demographic data were missing for some testing samples (see Table 1 and Appendix A for an overview of available data).

Table 1
States with different configurations of NAEP data available for analysis, 1992–2002

	Reading						Writing		
Year	1992	1994	1998	1998	2002	2002	1998	2002	2002
Grade	4	4	4	8	4	8	8	4	8
Missing Data	$ Edu.	$ Edu.	Edu.		Edu.			Edu.	
Total States	42	40	40	37	44	42	36	44	42
	Math						Science		
Year	1992	1992	1996	1996	2000	2000	1996	2000	2000
Grade	4	8	4	8	4	8	8	4	8
Missing Data	$ Edu.	$ Edu.	Edu.		Edu.			Edu.	
Total States	42	42	44	41	41	40	41	39	38

$ = family income; Edu. = parent education

The relationships between high-stakes testing policies and student achievement were assessed in three ways: relationships to single-year achievement, relationships to changes in achievement between testing periods within each grade level (fourth grade and eighth grade), and relationships to changes in achievement between testing periods within a cohort (fourth to eighth grade), using the following research questions.

Relationships to single year achievement. Do achievement scores differ between states with and without high stakes testing; do the demographic characteristics of the testing samples predict their achievement; and with the demographic characteristics controlled, do indicators of high-stakes testing policies predict the achievement of testing samples?

Relationships between testing periods within grade level. Do changes in achievement scores within grade level from one testing period to the testing period four years later (e.g., changes in fourth grade testing samples from one testing period to fourth grade testing samples four years later) differ between states with and without high stakes testing; do the changes in demographic characteristics of the testing samples predict these changes in achievement between two testing periods; and with the demographic characteristics controlled, do indicators of high-stakes testing policies predict changes in achievement between testing periods (within each grade level)?

Relationships between testing periods within a cohort. Do changes in achievement scores of cohort testing samples from their fourth grade testing period to their eighth grade testing period differ between states with and without high stakes testing; do the changes in demographic characteristics of cohort testing samples predict changes in achievement from fourth grade to eighth grade; and with the demographic characteristics controlled, do indicators of high-stakes testing policies predict changes in achievement of cohort testing samples from fourth to eighth grade?

Method

Units of Analysis

The units of analysis in this study were the testing samples of students taking the NAEP. A testing sample was defined as the group of students in one grade level (fourth grade or eighth grade) who took the NAEP at any given testing period (e.g., 1992, 1996, 2000) in one state (there were 50 states including DC, excluding *SD*). Therefore, any one state could have a total of 6 possible testing samples (fourth and eighth grade samples at each of three testing periods), for a total of 300 (50 x 6) possible testing samples (see Appendix A for a complete display of available testing samples), each with a unique set of demographic characteristics and unique achievement score. However, it is important to note that not every state administered the test at each testing period or to both grade levels, and many states were missing demographic data for some testing samples; therefore the number of "participants" varied among testing periods (see Table 1 and Appendix A). In addition, the writing and science subject tests were administered at only two testing periods, so there were fewer testing samples for those subjects. The number of valid testing samples is reported in the results section for each analysis.

Measures

NAEP achievement. Three indicators of NAEP achievement were used in this study. First, scale scores and changes in scale scores were used as the primary indicators of achievement. Scale scores were used as they represent average levels of achievement and gains in achievement across all students within the testing samples. Given the wide variability, scale scores would likely be most sensitive to any difference or changes in achievement that might exist.

Second, the percentage of students reaching *proficiency* and changes in the percentage of students reaching *proficiency* across testing years were used as another indicator of NAEP achievement. The major goal of NCLB is for 100% of students to reach *proficiency* on their state achievement tests by academic year 2014–2015. The federal legislation requires that schools and states report the percentage of students who reach Proficient, with that statistic tied to sanctions when federally mandated goals are not met. Similarly, in the current study, the percentage of students reaching Proficient (and changes in these percentages across testing years) on the NAEP was of interest (and likewise used in previous studies—see Carnoy & Loeb, 2002; Nichols et al., 2006). Unfortunately, the cut scores for reaching levels of Proficient may not necessarily match the levels of achievement specified as Proficient across states. Indeed, there is some indication that the level of Proficient on the NAEP is set higher than what most states would regard as Proficient (Fuller, Gesicki, Kang, & Wright, 2006; Linn, 2000) and its validity has not been tested against actual student knowledge (Linn, 2000). Also, states often change their own cut-scores over time, making comparisons across states using any assessment difficult (Linn, 2001). In fact, some states have lowered their cut-scores so that they can meet the requirements set by NCLB to reach 100% proficiency. Unfortunately, this practice does not allow states to differentiate between gains of lower performing and higher performing students because a greater range of students is being "shifted" into the Proficient levels. Therefore, although Proficient on the NAEP may not be aligned with state levels of Proficient, using percentage of students reaching Proficient on the NAEP as one of the achievement indicators allowed us to assess the impact of high-stakes testing policies on the higher performing students (a group often ignored when assessing growth in achievement).

Nevertheless, being mindful of the limitations of using percent Proficient as one of the achievement indicators, we chose to include percentage of students reaching Basic as our third indicator of achievement. It has been argued that the Basic level on the NAEP may be more closely aligned to levels of Proficient across states (Fuller et al., 2006). Using percentage of students reaching Basic also allowed us to examine the changes in the achievement of lower performing students. This way we could compare the differential effects of high-stakes testing on lower and higher performing students.

Demographic characteristics. Survey and school records data were collected at the time of each NAEP testing. Demographic data for each testing sample included one or more of the following characteristics (see Table 1 for missing variables): low family income (percentage of test-takers eligible for the federal free and reduced lunch program), parent education (percentage of test-takers with a parent with a college degree), ethnicity (percentage of Black and Hispanic test-takers), and excluded test-takers (percent of disabled and limited English proficient [LEP] test-takers excluded from testing).

High-stakes testing policy indicators. Three indicators of high-stakes testing policies were derived from two former studies. In their study, Amrein and Berliner (2002b) identified when each state initiated high-stakes testing policies. Those dates were used to establish whether a high-stakes testing policy was in place at the time of each NAEP testing. Therefore, in the current study, each testing sample was identified as to whether or not it belonged in the high-stakes group. That date also was used to establish, as a second indicator, the number of years high-stakes testing had been in place for each testing sample. The third high-stakes testing policy indicator was the high-stakes index developed by Carnoy and Loeb (2002) in their study using the NAEP. This strength of accountability indicator is measured on a scale of 0 to 5 and is based on several factors present before 2001 accountability mandates were in place. According to the scale's creators, it "captures degrees of state external pressure on schools to improve student achievement according to state-defined performance criteria" (Carnoy & Loeb, 2002, p. 311). The factors to determine each states' accountability index included whether or not states tested at the elementary and middle school levels, whether or not test results were reported to the state, whether or not scores were subject to sanctions or rewards (and the strength of such sanctions), and whether or not a high school exit exam was present.

Analysis Procedures

The purpose of this study was to determine if states' high-stakes testing policies contributed significantly to students' NAEP achievement beyond what could be predicted based on the demographic characteristics of test-takers. Multiple regressions were used to examine the relationships of the indicators of high-stakes testing policies to each of the indicators of NAEP achievement, after controlling for demographic characteristics. Results are reported separately for each subject area of the NAEP.

Results

Reading

Single year achievement. Means and standard deviations for the three indicators of students' reading achievement (NAEP scale score, percentage of students reaching Basic, and percentage of students reaching Proficient) and the demographic characteristics of all available testing samples are shown in Table 2, separated by whether or not the testing samples came from states identified as high-stakes at the time of the testing. There were 245 valid testing samples (94 samples in states with high stakes testing and 151 in states without high stakes testing); however, family income was not available for the 1992 or 1994 testing periods, and parent education was not available for any of the fourth grade testing samples. A listwise deletion of testing samples with missing data within each analysis produced smaller samples sizes for the analyses that included family income and/or parent education (slight variations in samples sizes across analyses are due to single cases of missing data within one or two states' data). Sample sizes are shown in the tables for each analysis.

Results revealed that across all testing samples, there were no differences between those with high stakes testing and those without high stakes testing on either reading scale scores or percentage of students reaching Basic on the reading achievement portion of the NAEP; however, those with high-stakes testing policies in place had a lower proportion of students reaching Proficient than did those without high-stakes testing. The demographic characteristics of the testing samples with and without high-stakes testing also differed significantly: in high stakes states, the percentage of families with low income was higher, the percentage of parents with a college education was lower, the percentage Black and Hispanic students was higher, and the percentage of students excluded from testing was higher (see Table 2). Indeed, bivariate correlations between the demographics and reading scale scores revealed that all four demographic factors were significantly associated with a higher scores: lower percentage of families with low income (n = 163, r = -.54, p < .001), higher percentage of parents with a college education (n = 78, r = .65, p < .001), lower percentage of Black and Hispanic students (n = 245, r = -.23, p < .001), and lower percentage of disabled and LEP students excluded from testing (n = 245, r = -.23, p < .01).

Table 2
Means and Standard Deviations of Single Year Reading Achievement and Demographics

Variable	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	—	151	227.79	22.82	94	233.26	24.08	-1.78
% Above Basic	.81***	151	64.72	11.62	94	64.41	10.02	0.21
% Above Proficient	.34***	151	29.91	9.53	94	26.98	6.11	2.93**
% Low family income	-.54***	85	37.27	12.38	78	43.13	11.29	-3.15**
% Parent college educ	.65***	39	45.74	6.51	39	41.62	5.67	2.99**
% Black/Hispanic	-.23***	151	21.21	19.74	94	30.66	14.78	-4.00***
% Excluded disabled	-.23**	151	5.99	2.18	94	7.15	2.42	-3.93***

* p < .05, ** p < .01, *** p < .001

In the regression analyses (see Table 3), the four demographic factors predicted almost 70% of the variance among testing samples on all three indicators of reading achievement (68% of scale scores, 65% of percentage reaching Basic, and 72% of percentage reaching Proficient), with each demographic factor making a unique contribution to the prediction (Beta coefficients are reported in Table 3). In the second step of the regression, the high-stakes testing indicators did not add to the prediction of any of the reading achievement factors.

In addition, because parent education data were not available for fourth grade testing samples, a second regression was run without parent education in the equation, thereby effectively doubling the sample size (see Table 3). Without parent education, the demographic factors now accounted for 35% of the variance in scale scores, 53% of the variance in percentage reaching Basic and 48% of the variance percentage reaching Proficient. Beyond demographics, the high stakes testing variables now accounted for 11% of the variance in scale scores and 9% of the variance in percentage of students reaching Basic, but no additional variance in percentage of students reaching Proficient. In fact, the beta coefficients revealed that samples having high stakes testing had lower reading scores and fewer students reaching Basic (as indicated by the negative coefficient). However, in high stakes samples, years of testing and the high-stakes index predicted higher achievement. In general, these analyses indicate that demographics are considerably more significant in predicting students' reading achievement than indicators of high stakes testing; in fact, the effects of high stakes testing on scale scores and reaching Basic disappears when parent education is present in the equation.

Table 3
Predicting Three Indicators of Single Year Reading Achievement

	Scale Score		% Above Basic		% Above Proficient	
	parent edu	*no parent edu*	*parent edu*	*no parent edu*	*parent edu*	*no parent edu*
Variable	*n* = 77	*n* = 162	*n* = 77	*n* = 162	*n* = 77	*n* = 162
Demographics	R^2 =.68***	R^2 =.35***	R^2 =.65***	R^2 =.53***	R^2 =.72***	R^2 =.48***
% Low income	-0.21*	-0.62***	-0.17	-0.65***	-0.30**	-0.52***
% Parent college	0.50***	—	0.42***	—	0.49***	—
% Black/Hispanic	-0.38***	0.27**	-0.47***	-0.07	-0.31***	-0.27***
% Excl/disabled	0.28***	-0.24**	0.27*	-0.11	0.20**	0.05
High Stakes Indicator	R^2 =.03	R^2 =.11***	R^2 =.03	R^2 =.09***	R^2 =.03	R^2 =.02
Present (yes/no)	-0.06	-0.18	-0.06	-0.24**	-0.09	-0.20
Years w/ testing	0.20	0.39***	0.22	0.43***	0.13	0.19
High-Stakes Index	0.11	0.23**	0.09	0.17*	0.20	0.12

* $p < .05$, ** $p < .01$, *** $p < .001$. Coefficients are standardized beta weights.

Four-year change in reading achievement within grade level. Means and standard deviations for changes in reading achievement and the changes in demographic characteristics between four year testing periods within the same grade levels (fourth to fourth grade and eighth to eighth grade) are shown in Table 4, separated by whether or not high-stakes testing policies were present during the change period. There were 106 valid testing samples: 35 states participated in fourth grade reading testing in both 1994 and 1998, 37 states participated in fourth grade reading testing in both 1998 and 2002, and 34 states participated in eighth grade reading testing in both 1998 and 2002. Results revealed that high stakes samples demonstrated slightly greater gains in scale scores (less than two points) and in percentage of students reaching Basic (less than 2%), but no gains in proportion of students reaching Proficient reading levels. In addition, high stakes samples had a slightly greater increase in percentage of low-income families (only 2%). There were no significant bivariate correlations between changes in demographic factors and changes in reading achievement between testing periods (see Table 4).

Table 4
Means and Standard Deviations of Change in Reading Achievement (Same Grade)

Variable	Correlation with scale score	Not High-Stakes			High-Stakes			t
		n	Mean	SD	n	Mean	SD	
NAEP Scale score	.98***	53	0.32	4.02	53	1.97	4.55	-2.30*
% Above Basic	.93***	53	0.15	3.92	53	1.89	5.33	-2.26*
% Above Proficient	.87***	53	0.74	2.80	53	1.83	3.51	-1.78
% Low family income	-.25	31	2.16	4.33	40	4.16	3.73	-2.09*
% Parent college educ	.31	13	3.31	2.98	20	3.50	2.50	-0.20
% Black/Hispanic	.08	53	1.32	2.85	53	1.92	4.04	-0.89
% Excluded disabled	.16	53	-0.41	2.14	53	-0.20	2.99	-0.43

* $p < .05$, ** $p < .01$, *** $p < .001$

In the regression analyses (see Table 5), the demographic characteristics together predicted no significant change in reading scale scores, or in percentage reaching Basic, but predicted 33% of the change in students reaching Proficient within grade levels ($R^2 = .33, p < .05$), with parent education and percentage of disabled and LEP students excluded from the testing contributing significant amounts of unique variance (Beta coefficients are shown in Table 3). The high-stakes testing indicators did not add significantly to the prediction of changes in any indicators of reading achievement. In a second regression, excluding parent education (effectively doubling the sample size by including fourth grade testing samples), the changes in demographics significantly accounted for 20%, 24% and 18% of the variance in changes in reading scale scores, proportion of students reaching Basic, and proportion of students reaching Proficient, respectively; but high stakes testing indicators predicted no additional variance.

It should be noted that when predicting change scores, the proportions of variance predicted will be smaller than when predicting single year scores due to restrictions of range and variability; thus R^2 coefficients were around 20–30% in these analyses, rather than the 60–70% seen in the previous analyses. Similarly, we felt compelled to note that although the high stakes indicators were predicting as much as 9% of additional variance in changes in reading achievement, they were not significant because of the lower number of samples in the analyses. We were especially intrigued by the pattern of relationships revealed by the Beta coefficients. Whereas having high stakes testing was indicative of greater changes in reading achievement, more years of testing and higher index scores predicted smaller changes in achievement. Although these patterns are interesting, it was the demographic factors that continued to account for a significantly greater proportion of the variance in changes in reading achievement than any indicator of high stakes testing.

Table 5
Predicting Change in Three Indicators of Reading Achievement (Same Grade)

	Scale Score		% Above Basic		% Above Proficient	
	parent edu *n* = 77	*no parent edu* *n* = 162	*parent edu* *n* = 77	*no parent edu* *n* = 162	*parent edu* *n* = 77	*no parent edu* *n* = 162
Demographics	R^2 = .26	R^2 = .20**	R^2 = .29	R^2 = .24*	R^2 = .33*	R^2 = .18**
% Low income	-0.07	-0.17	-0.15	-0.08	-0.04	-0.17
% Parent college	0.27	—	0.11	—	0.33*	—
% Black/Hispanic	-0.03	0.34**	-0.07	-0.07	0.13	0.27*
% Excl/disabled	0.40	0.41**	0.45	0.44*	0.48**	0.40**
High Stakes Indicator	R^2 = .09	R^2 = .06	R^2 = .05	R^2 = .07	R^2 = .09	R^2 = .09
Present (yes/no)	0.52	0.41	0.23	0.27	0.52	0.51
Years w/ testing	-0.30	-0.39	0.03	0.01	-0.28	-0.42
High-Stakes Index	-0.27	-0.01	-0.03	-0.01	-0.19	-0.03

* p < .05, ** p < .01, *** p < .001. Coefficients are standardized beta weights.

Four-year change in reading achievement within cohort. Means and standard deviations for the change in reading achievement and the changes in demographic characteristics between four year testing periods within the same cohort (fourth to eighth grade) are shown in Table 6, separated by whether or not high-stakes testing policies were present during the change period. There were 68 valid testing samples: 33 states participated in fourth to eighth grade reading testing from 1994 to 1998 and 35 states participated in fourth to eighth grade reading testing from 1998 to 2002. (Note that parent education data were not available within cohorts because those data were not collected in fourth grade testing samples.) High stakes samples had a slightly greater increase in proportion of students reaching Basic (less than 3%) and a slightly smaller decrease in the percentage of low-income families than did samples without high stakes testing (see Table 6).

Table 6
Means and Standard Deviations of Change in Reading Achievement (Cohort)

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	.94***	34	48.59	4.90	34	49.68	3.72	-1.03
% Above Basic	.90***	34	14.12	4.69	34	16.56	4.49	-2.19*
% Above Proficient	.68***	34	1.88	2.96	34	2.53	3.10	-0.88
% Low family income	-.45**	14	-7.59	5.16	21	-4.48	3.01	-2.26*
% Parent college educ	—	—	—	—	—	—	—	—
% Black/Hispanic	-.19	34	2.18	3.14	34	1.24	3.77	1.12
% Excluded disabled	.22	34	-1.59	2.25	34	-1.50	3.21	-0.14

* p < .05, ** p < .01, *** p < .001

In the regression analyses, the demographic characteristics significantly predicted changes in percentage of students reaching Basic (27% of the variance) and percentage of students reaching Proficient (28% of the variance) from fourth to eighth grade (see Table 7), but the high-stakes testing indicators did not add to the prediction of any of the achievement indicators. This regression, however, included only the 1998–2002 cohort, because parent income data were not available in the 1994 fourth grade sample. A second regression was run without family income, thereby effectively doubling the sample size (including both cohorts, 1994–1998 and 1998–2002). The demographic factors continued to predict a significant proportion of the variance in the change in percentage of students reaching Basic (R^2 = .14, p < .01) and students reaching Proficient (R^2 = .12, p < .05). In addition, the high stakes indicators predicted 16% of additional variance in change in percentage reaching Basic (see Table 7), with the high stakes indicators positively predicting greater changes in achievement within cohort. Unfortunately, due to missing data, it is impossible to know whether or not the high stakes indicators would have remained significant had we been able to control from parent education. However, trends from other analyses would suggest that the addition of parent education would weaken the effects of high stakes testing.

Table 7
Predicting Change in Three Indicators of Reading Achievement (Cohort Group)

	Scale Score		% Above Basic		% Above Proficient	
	income *n* = 34	*no income* *n* = 67	*income* *n* = 34	*no income* *n* = 67	*income* *n* = 34	*no income* *n* = 67
Demographics	R^2 = .20	R^2 = .07	R^2 = .27*	R^2 = .14**	R^2 = .28*	R^2 = .12*
% Low income	-0.19	—	-0.15	—	-0.18	—
% Parent college	—	—	—	—	—	—
% Black/Hispanic	-0.09	-0.11	-0.08	-0.15	-0.10	-0.30*
% Excl/disabled	0.29	0.20	0.41	0.28*	0.37	0.09
High Stakes Indicator	R^2 = .13	R^2 = .10	R^2 = .17	R^2 = .16**	R^2 = .12	R^2 = .02
Present (yes/no)	0.42	0.07	0.33	0.04	0.50	0.09
Years w/ testing	-0.28	-0.04	-0.10	0.12	-0.47	-0.13
High-Stakes Index	0.18	0.30	0.22	0.31*	0.12	0.14

* p < .05, ** p < .01, *** p < .001. Coefficients are standardized beta weights.

Writing

Single year achievement. Means and standard deviations for the three indicators of writing achievement and the demographic characteristics of all available testing samples are shown in Table 8, separated by whether or not the testing sample came from a state identified as high-stakes at the time of the testing. There were 122 valid testing samples: eighth grade in 1998 and 2002 and fourth grade in 2002 (see Table 1, note there was no cohort sample). Also, the absence of parent education data from fourth grade testing samples reduced the number of samples available in analyses that included parent education.

Table 8
Means and Standard Deviations of Single Year Writing Achievement

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	—	59	149.09	8.69	63	149.22	7.63	-0.09
% Above Basic	.94***	59	82.79	5.81	63	83.70	4.50	-0.96
% Above Proficient	.95***	59	25.47	7.89	63	24.44	7.69	0.73
% Low family income	-.53***	59	36.61	12.17	63	43.12	11.75	-3.00**
% Parent college educ	.52***	38	48.92	5.91	39	44.05	5.42	3.76***
% Black/Hispanic	-.40***	59	19.92	21.29	63	31.30	14.45	-3.48***
% Excluded disabled	-.06	59	3.95	1.36	63	4.79	1.71	-3.07**

* $p < .05$, ** $p < .01$, *** $p < .001$

Analyses showed that writing proficiency did not differ between states with and without high stakes testing (see Table 8). Bivariate correlations between demographics and writing achievement revealed three demographic characteristics that were significantly associated with higher writing achievement: fewer low income families, higher parent education, and fewer minority students.

Table 9
Predicting Three Indicators of Single Year Writing Achievement

	Scale Score		% Above Basic		% Above Proficient	
	parent edu *n* = 76	*no parent edu* *n* = 121	*parent edu* *n* = 76	*no parent edu* *n* = 121	*parent edu* *n* = 76	*no parent edu* *n* = 121
Demographics	R^2 = .33***	R^2 = .25***	R^2 = .27***	R^2 = .22***	R^2 = .32***	R^2 = .31***
% Low income	-0.44**	-0.46***	-0.32	-0.39***	-0.48**	-0.61***
% Parent college	0.23	—	0.26	—	0.19	—
% Black/Hispanic	0.03	-0.06	-0.03	-0.13	0.08	0.08
% Excl/disabled	0.11	0.27**	0.15	0.26**	0.06	0.18*
High Stakes Indicator	R^2 = .07	R^2 = .01	R^2 = .12**	R^2 = .02	R^2 = .04	R^2 = .01
Present (yes/no)	-0.02	0.08	0.13	0.18	-0.13	-0.03
Years w/ testing	0.28	-0.04	0.26	-0.02	0.29	0.01
High-Stakes Index	0.12	0.05	0.12	0.01	0.12	0.14

* $p < .05$, ** $p < .01$, *** $p < .001$. Coefficients are standardized beta weights.

In the regression analyses (see Table 9), the demographic factors predicted on average 30% of the variance among states' achievement on the NAEP writing test (see Table 9), with parent income making the greatest unique contribution to the prediction (Beta coefficients are reported in Table 9). The high stakes indicators significantly added to the prediction of a greater proportion of students reaching Basic ($R^2 = .12$, $p < .01$). When excluding parent education from the regression equation, the demographics continued to predict significant proportions of variance in writing achievement; however, the high-stakes testing indicators no longer added to the prediction of writing achievement. On the writing test, similar to the results in reading, demographics continue to predict the greater proportion of variance in students' achievement on the NAEP.

Four-year change in writing achievement within grade level. Means and standard deviations for the changes in writing achievement and in demographic characteristics between four year testing periods within the same grade levels (eighth to eighth grade) are shown in Table 10. In these analyses, there were only 33 testing samples (states that participated in eighth grade writing testing in both 1998 and 2002). Changes in writing achievement within grade level did not differ between states with and without high stakes testing. Bivariate correlations revealed positive relationships between changes in writing scores and both parent education and percentage of students excluded from testing. In the regression analyses (see Table 11), changes in demographic factors significantly predicted changes in writing scale scores and in proportion of students reaching Basic and Proficient within grade levels; and high-stakes testing indicators did not add significantly to the prediction of changes in any of the indicators of writing achievement.

Table 10
Means and Standard Deviations of Change in Writing Achievement (Same Grade)

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	.84***	13	2.54	2.70	20	4.25	5.15	-1.10
% Above Basic	.89***	13	0.54	2.60	20	1.60	4.33	-0.79
% Above Proficient	.90***	13	3.38	2.84	20	5.35	4.58	-1.38
% Low family income	-.30	13	-2.93	4.63	20	-4.27	4.71	-0.80
% Parent college educ	.44**	13	-2.62	1.19	20	-2.80	2.71	0.23
% Black/Hispanic	-.32	13	2.65	2.20	20	2.20	2.46	0.49
% Excluded disabled	.45**	13	-0.94	1.24	20	-0.55	2.21	-0.63

* $p < .05$, ** $p < .01$, *** $p < .001$

Four-year change in writing achievement within cohort. The necessary data for these analyses were not available, as fourth graders were not given the writing test in 1998.

Table 11
Predicting Change in Three Indicators of Writing Achievement (Same Grade)

	Scale Score *n* = 32	% Above Basic *n* = 32	% Above Proficient *n* = 32
Demographics	R^2 = .37*	R^2 = .30*	R^2 = .32*
% Low income	-0.11	-0.07	-0.19
% Parent college	0.31	0.24	0.39*
% Black/Hispanic	-0.20	-0.30	-0.07
% Excl/disabled	0.26	0.20	0.12
High-stakes Indicatoı	R^2 = .04	R^2 = .05	R^2 = .04
Present (yes/no)	0.35	0.29	0.29
Years w/ testing	-0.18	-0.03	-0.12
High-Stakes Index	-0.23	-0.26	-0.02

* *p* < .05, ** *p* < .01, *** *p* < .001. Coefficients are standardized beta weights.

Mathematics

Single year achievement. Means and standard deviations for the achievement indicators in math and the demographic characteristics of all available testing samples are shown in Table 12. There were 250 valid testing samples (173 samples in states with high stakes testing and 77 in states without high stakes testing); however family income was not available for the 1992 testing period, and parent education was not available for fourth grade testing samples. A listwise deletion of testing samples with missing data within each analysis produced smaller samples sizes for analyses that included family income and/or parent education (slight variations in samples sizes across analyses are due to single cases of missing data within one or two states' data). Sample sizes are shown in the tables for each analysis.

Table 12
Means and Standard Deviations of Single Year Math Achievement

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	—	173	245.35	26.53	77	243.90	24.32	0.41
% Above Basic	.31***	173	61.02	12.38	77	58.97	10.21	1.27
% Above Proficient	.48***	173	20.45	7.08	77	18.64	6.59	1.91
% Low family income	-.60***	101	34.88	11.73	65	40.87	12.29	-3.16**
% Parent college educ	.73***	83	42.29	6.75	38	38.18	5.80	3.24**
% Black/Hispanic	-.31***	173	20.67	19.56	77	28.97	15.63	-3.28***
% Excluded disabled	.14*	173	6.32	2.44	77	7.91	2.62	-4.68***

* *p* < .05, ** *p* < .01, *** *p* < .001

Results revealed that math achievement did not differ between states with and without high stakes testing (see Table 12). However, bivariate correlations between demographics and math scale scores revealed all four demographic characteristics were significantly associated with higher math achievement: fewer low income families, higher parent education, fewer minorities, and more students excluded from testing.

In the regression analyses (see Table 13), demographic factors predicted about 77% of the variance among the states' achievement on the NAEP math test (78% in scale scores, 77% for students reaching Basic, and 76% for students reaching Proficient), with higher parent education, fewer low income, and less ethnicity making unique contributions to the prediction (Beta coefficients are reported in Table 13). In the second step of the regression, the high-stakes index predicted a small (3%), but significant, proportion of additional variance in scale scores and in percentage of students reaching Basic. However, similar to the results with reading achievement, beta coefficients revealed that high stakes samples had lower achievement than did non-high stakes samples; but within high stakes samples, more years of testing and higher index scores predicted greater achievement.

The regression was run again excluding parent education (so that fourth grade testing samples were included in the analysis), and the demographic factors continued to predict significant proportions of variance in math achievement (38% of scale scores, 57% of students reaching Basic, and 63% of students reaching Proficient) with family income and ethnicity predicting the most unique variance. The high stakes indicators continued to predict a small, but significant, additional proportion of variance in percentage of students reaching Basic; but again, having high stakes was related to lower achievement, although years of testing and the high stakes indicator predicted higher math achievement.

Table 13
Predicting Three Indicators of Single Year Math Achievement

	Scale Score		% Above Basic		% Above Proficient	
	parent edu *n* = 77	*no parent edu* *n* = 162	*parent edu* *n* = 77	*no parent edu* *n* = 162	*parent edu* *n* = 77	*no parent edu* *n* = 162
Demographics	R^2 = .78***	R^2 = .38***	R^2 = .77***	R^2 = .57***	R^2 = .76***	R^2 = .63***
% Low income	-0.38***	-0.72***	-0.32**	-0.45***	-0.38***	-0.70***
% Parent college	0.40***	—	0.39***	—	0.48***	—
% Black/Hispanic	-0.67**	0.26***	-0.33***	-0.42***	-0.16	-0.16*
% Excl/disabled	0.01	-0.13	0.02	0.16**	0.03	0.10
High Stakes Indicator	R^2 = .03*	R^2 = .04*	R^2 = .03*	R^2 = .01	R^2 = .02	R^2 = .01
Present (yes/no)	-0.11	-0.07	-0.12	-0.09	-0.06	-0.07
Years w/ testing	0.19*	0.18	0.21*	0.12	0.09	0.05
High-Stakes Index	0.18*	0.17*	0.16*	0.07	0.18	0.15

* $p < .05$, ** $p < .01$, *** $p < .001$. Coefficients are standardized beta weights.

Four-year change in math achievement within grade level. Means and standard deviations for the change in the three indicators of math achievement and the changes in demographic characteristics between four year testing periods within the same grade levels (fourth to fourth grade and eighth to eighth grade) are shown in Table 14. There were 144 valid testing samples: 38 states participated in fourth grade math testing in both 1992 and 1996, 36 participated in fourth grade math testing in both 1996 and 2000, 36 states participated in eighth grade math testing in both 1992 and 1996, and 34 participated in eighth grade math testing in both 1996 and 2000. Analyses showed that high stakes samples had slightly greater change in scale scores (less than 2 points) and slightly greater change in proportion of students reaching Basic (less than 3%); but changes in proportion of students reaching Proficient did not differ between states with and without high stakes testing. Bivariate correlations between demographics and math achievement revealed that a decrease in the percentage of Black and Hispanic students was related to increases in math achievement.

Table 14
Means and Standard Deviations of Change in Math Achievement (Same Grade)

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	.99***	84	2.31	3.27	60	3.60	3.07	-2.39*
% Above Basic	.86***	84	2.39	3.47	60	4.87	3.75	-4.07*
% Above Proficient	.70***	84	2.11	2.56	60	2.90	2.63	-1.81
% Low family income	-.18	33	2.14	3.33	37	1.13	2.85	1.37
% Parent college educ	.17	41	1.41	2.25	29	2.28	2.43	-1.53
% Black/Hispanic	-.23**	84	1.44	2.55	60	1.57	3.44	-0.25
% Excluded disabled	.04	84	1.66	1.78	60	1.79	2.45	-0.37

* $p < .05$, ** $p < .01$, *** $p < .001$

However, in the regression analyses (see Table 15), the combined demographic factors did not significantly predict changes in math achievement; but the high stakes index significantly added to the prediction of changes in math scale scores, with greater high stakes index scores predicting higher math scale scores (both the presence of high stakes testing and years of testing predicted scale scores in the negative direction). In the second regression, with parent education removed (thereby including fourth graders and doubling the sample size), changes in demographics still did not predict changes in math achievement. The high stakes indicators continued to add to the prediction of scale scores, with the presence of high stakes testing and years of testing negatively related to gains in math achievement, and the high stakes indicator positively related to changes in math scale scores. The high stakes indicators still did not predict changes in either percent Basic or percent Proficient.

Table 15
Predicting Change in Three Indicators of Math Achievement (Same Grade)

	Scale Score		% Above Basic		% Above Proficient	
	parent edu	*no parent edu*	*parent edu*	*no parent edu*	*parent edu*	*no parent edu*
	n = 32	*n* = 70	*n* = 32	*n* = 70	*n* = 32	*n* = 70
Demographics	R^2 = .17	R^2 = .07	R^2 = .20	R^2 = .19	R^2 = .15	R^2 = .12
% Low income	-0.20	-0.22	-0.37	-0.37	-0.26	-0.30
% Parent college	0.13	—	-0.07	—	0.00	—
% Black/Hispanic	-0.21	-0.05	-0.17	-0.16	-0.01	-0.03
% Excl/disabled	0.07	0.07	0.08	-0.11	0.25	0.15
High Stakes Indicator	R^2 = .25*	R^2 = .15*	R^2 = .13	R^2 = .13	R^2 = .17	R^2 = .05
Present (yes/no)	-0.12	0.18	-0.34	-0.34	-0.29	0.13
Years w/ testing	-0.11	-0.16	-0.25	-0.22	-0.04	-0.18
High-Stakes Index	0.66*	0.35*	0.48	0.45	0.61	0.21

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Coefficients are standardized beta weights.

Four-year change in mathematics achievement within cohort. Means and standard deviations for the changes in math achievement and in demographic characteristics between four year testing periods within the same cohort (fourth to eighth grade) are shown in Table 16. There were 74 valid testing samples: 39 states participated in fourth to eighth grade math testing from 1992 to 1996 and 35 states participated in fourth to eighth grade reading testing from 1996 to 2000. In the analyses on math achievement, the states with high stakes testing had a slightly smaller change in math scale scores (less than 2 points), but a slightly higher change in proportion of students reaching Proficient (less than 1%) than did states with no high stakes testing.

Table 16
Means and Standard Deviations of Change in Math Achievement (Cohort)

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	.94***	43	51.39	3.92	31	49.63	3.35	1.98*
% Above Basic	.89***	43	2.29	3.57	31	1.20	3.21	1.33
% Above Proficient	.88***	43	0.56	2.39	31	1.19	2.24	2.02*
% Low family income	-.19	16	-5.62	3.23	19	-7.30	2.68	1.65
% Parent college educ	—	—	—	—	—	—	—	—
% Black/Hispanic	-.05	43	0.49	2.26	31	-.03	3.16	0.77
% Excluded disabled	.11	43	0.70	1.92	31	1.91	2.45	-1.73

* $p < .05$, ** $p < .01$, *** $p < .001$

Furthermore, in the regression analysis, changes in demographic characteristics did not significantly predict changes in any indicators of math achievement from fourth to eighth grade (see

Table 17; note that parent education was not available for either cohort and that family income was available for only one cohort); and the high stakes indicators predicted no additional variance in any of the math achievement factors. It was interesting to note, however, that although not significant, the proportions of variance predicted by the high stakes indicators were fairly large (in the 15–20% range) in the equation with fewer samples; but as the number of samples increased, the proportion of variance predicted by the high stakes factors decreased. In addition, the direction of the beta coefficients did not favor the presence of high stakes testing nor the number of years of testing in producing higher levels of math achievement.

Table 17
Predicting Change in Math Achievement (Cohort)

	Scale Score		% Above Basic		% Above Proficient	
	income *n* = 34	*no income* *n* = 67	*income* *n* = 34	*no income* *n* = 67	*income* *n* = 34	*no income* *n* = 67
Demographics	R^2 = .05	R^2 = .02	R^2 = .16	R^2 = .05	R^2 = .03	R^2 = .02
% Low income	-0.16	—	-0.25	—	-0.13	—
% Parent college	—	—	—	—	—	—
% Black/Hispanic	-0.07	-0.05	-0.13	-0.14	0.00	-0.02
% Excl/disabled	-0.13	-0.14	-0.25	-0.18	-0.10	-0.14
High Stakes Indicator	R^2 = .17	R^2 = .09	R^2 = .15	R^2 = .05	R^2 = .21	R^2 = .09
Present (yes/no)	-0.39	-0.17	-0.35	-0.14	-0.24	-0.01
Years w/ testing	-0.24	-0.19	-0.24	-0.16	-0.42	-0.28
High-Stakes Index	0.31	0.05	0.51	0.18	0.24	-0.04

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Coefficients are standardized beta weights.

Science

Single year achievement. Means and standard deviations for science achievement and the demographic characteristics of all available testing samples are shown in Table 18. There were 118 valid testing samples: eighth grade in 1996 and 2000 and fourth grade in 2000 (see Table 1—note the absence of cohort analysis). Results showed that states with no high stakes testing had higher science achievement on all indicators, but they also had fewer low income families, higher parent education, a lower percentage of Black and Hispanic students, and fewer students excluded from testing. Bivariate correlations between demographics and science achievement revealed that these demographic characteristics were significantly associated with science achievement.

Table 18
Means and Standard Deviations of Single Year Science Achievement

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	—	65	151.19	9.36	53	146.27	7.73	3.03**
% Above Basic	.95***	63	65.07	10.83	51	59.16	10.31	2.92**
% Above Proficient	.96***	65	30.28	8.14	53	25.53	7.41	3.31**
% Low family income	-.74***	65	31.92	8.90	53	39.69	12.22	3.99***
% Parent college educ	.66***	46	46.50	6.06	32	41.69	5.38	3.61***
% Black/Hispanic	-.81***	65	20.62	15.50	53	32.11	14.66	-4.11***
% Excluded disabled	.14	65	7.11	2.46	53	8.62	2.82	-3.05**

* $p < .05$, ** $p < .01$, *** $p < .001$

In the regression analyses (see Table 19), on average the four demographic factors predicted 79% of the variance among the states' achievement on the NAEP science test with parent education included in the equation, and only slightly less of the variance when parent education was excluded. In the second step of the regressions, the high-stakes testing indicators did not add to the prediction of any of the science achievement measures.

Table 19
Predicting Three Indicators of Single Year Science Achievement

	Scale Score		% Above Basic		% Above Proficient	
	parent edu	*no parent edu*	*parent edu*	*no parent edu*	*parent edu*	*no parent edu*
	n = 70	*n* = 110	*n* = 70	*n* = 110	*n* = 70	*n* = 110
Demographics	R^2 = .78***	R^2 = .69***	R^2 = .80***	R^2 = .60***	R^2 = .79***	R^2 = .66***
% Low income	-0.24*	-0.37***	-0.24*	-0.18*	-0.21***	-0.45***
% Parent college	0.30***	—	0.30***	—	0.41***	—
% Black/Hispanic	-0.50***	-0.77***	-0.53***	-0.68***	-0.45***	-0.48***
% Excl/disabled	0.03	0.11	0.05	0.21**	0.10	0.15*
High Stakes Indicator	R^2 = .02	R^2 = .01	R^2 = .01	R^2 = .01	R^2 = .01	R^2 = .00
Present (yes/no)	-0.12	-0.11	-0.04	-0.04	-0.03	-0.03
Years w/ testing	0.18	0.14	0.14	0.09	0.12	0.04
High-Stakes Index	0.09	-0.01	0.03	-0.11	0.03	-0.04

* $p < .05$, ** $p < .01$, *** $p < .001$. Coefficients are standardized beta weights.

Four-year change in science achievement within grade level. Means and standard deviations for the changes in science achievement and the changes in demographic characteristics between four year testing periods within the same grade levels (fourth to fourth grade and eighth to eighth grade) are shown in Table 20. In these analyses, there were only 34 testing samples (states that participated in eighth grade science testing in both 1996 and 2000). Although the states with high stakes testing had a greater increase in proportion of students reaching Basic (see Table 20), the regression analyses revealed that the demographic factors predicted 30% to 40% of the variance in the change in science achievement (see Table 21); and, the high-stakes testing indicators did not add to the prediction.

Four-year change in science achievement within cohort. The necessary data for these analyses were not available as fourth graders were not given the science test in 1996.

Table 20
Means and Standard Deviations of Change In Science Achievement (Same Grade)

	Correlation with scale score	Not High-Stakes			High-Stakes			
		n	Mean	*SD*	*n*	Mean	*SD*	*t*
NAEP Scale score	.95***	16	0.07	1.98	18	1.56	3.26	-1.55
% Above Basic	.73***	16	-0.07	2.59	18	2.59	3.26	-2.48*
% Above Proficient	.73***	16	1.53	2.59	18	3.39	2.77	-1.97
% Low family income	-.18	16	-2.37	3.70	18	-0.85	3.03	1.30
% Parent college education	.52**	16	1.60	1.50	18	1.44	2.53	0.21
% Black/Hispanic	-.41*	16	1.73	2.19	18	0.78	3.41	0.94
% Excluded disabled/LEP	.05	16	1.88	1.36	18	2.26	2.75	-0.51

* p < .05, ** p < .01, *** p < .001

Table 21
Predicting Change in Three Indicators of Science Achievement (Same Grade)

	Scale Score *n* = 32	% Above Basic *n* = 32	% Above Proficient *n* = 32
Demographics	R^2 = .42**	R^2 = .32*	R^2 = .31*
% Low income	-0.11	-0.13	-0.07
% Parent college	0.50**	0.16	0.32
% Black/Hispanic	-0.38*	-0.50**	-0.38*
% Excl/disabled	-0.07	0.14	0.18
High-stakes Indicator	R^2 = .06	R^2 = .17	R^2 = .14
Present (yes/no)	0.14	0.52	0.37
Years w/ testing	0.19	0.11	0.21
High-Stakes Index	-0.08	-0.41	-0.35

* p < .05, ** p < .01, *** p < .001. Coefficients are standardized beta weights.

Discussion

The impact of characteristics (i.e., demographics) of the test-takers on aggregated test scores cannot be emphasized enough. When test scores are aggregated at any level (school, district, or state), the shared characteristics of the test-takers are strongly related to the variability in achievement results. Slight differences in demographics between states and slight changes in demographics over time can be associated with significant differences or changes in achievement. Marchant and Paulson (2001, 2005) previously demonstrated the need to consider demographics when interpreting aggregated SAT scores. Differences in test scores often are assumed to reflect differences in an intervention (such as the educational policies and practices of a state), when the scores reflect the collective nature of the test-takers (Marchant, 2005). In this study, simple comparisons of NAEP scores from states having high-stakes testing policies with those that do not revealed that high-stakes testing policies were often related to slightly lower NAEP scores (e.g., in reading and in science). However, simple comparisons of changes in NAEP scores over four years revealed that high-stakes testing often was related significantly to improved scores (e.g., in both reading and math). However, further analyses showed that most of these relationships (whether in favor of high stakes or non-high stakes states) disappeared once demographic differences were controlled. Indeed, past studies finding that high-stakes testing has relationships (either positive or negative) with achievement outcomes have not sufficiently controlled for demographics. Similarly, studies attributing significant achievement gains to high-stakes testing also have failed to adequately consider demographics.

In this study, the characteristics of the NAEP test-takers accounted for the majority of the variance among testing samples at the state level. These results confirmed the importance of controlling for ethnicity and disabled/LEP exclusions, as demonstrated in previous studies using NAEP data, but the inclusion of family income and parent education proved especially valuable in predicting variability among testing samples, with both of these factors consistently adding significantly to the prediction of achievement. In particular, in predicting single-year reading achievement, the high stakes indicators were found to add significantly to the prediction, but only when parent education was not controlled. In those testing samples in which parent education was collected and thereby controlled, the high stakes indicators no longer added significantly to the prediction of reading achievement. Similarly, the high stakes indicators added significantly to the prediction of changes in reading over four years within cohort, but only when family income and parent education data were not available to control. Indeed, when up to 70% of the variability among states' aggregated NAEP scores can be predicted by the average demographic characteristics of the states' test-takers—factors outside of the control of educational policies—educators and policy makers should be careful when attributing differences among states' performance to the policies alone. Likewise, when looking at changes in aggregated scores over time, it would be inappropriate to attribute those changes to educational policies or practices without careful consideration of other factors known to be associated with those scores.

Despite the large role that demographics played in predicting almost all of the achievement outcomes, there were several interesting patterns in the role of the high stakes indicators. In reading achievement, the high stakes indicators were found to add significantly to the prediction of scale scores and percentage of students reaching Basic, and to the prediction of cohort changes in percentage of students reaching Basic. Inspection of the Beta coefficients, however, showed that having high stakes was related to lower achievement outcomes; but years of high stakes testing and the high stakes index were related to higher achievement outcomes. Perhaps it could be argued that high stakes policies need to be in place longer or need to apply greater pressure in order to improve achievement. However, as discussed previously, these relationships disappeared when differences in all of the demographic characteristics among testing samples (i.e., family income and parent education) could be taken into account. Furthermore, high stakes indicators were not found to predict any variability in percentage of students reaching Proficient or in changes in percentage of students reaching Proficient over four years.

Similarly, in writing, the high stakes indicators did predict a greater percentage of students reaching Basic, even after controlling for demographics; but they did not account for any variability in students reaching Proficient. So, even though high stakes indicators may be able to have a small effect the writing performance of students at the Basic level, they showed no evidence of affecting the performance of higher achieving students. In science, the high stakes indicators had no impact on any of the indicators of performance. In general, we would argue that it is inappropriate to base a general school improvement reform on the use of high-stakes testing, calling into question the "No Child Left Behind" mandate requiring annual testing as a means to improve student learning.

The findings for math achievement proved to be a bit more complex. As with Carnoy and Loeb (2002) and Nichols et al. (2006), we acknowledge that differences and changes in math performance may be related to some aspect of high stakes testing policies, at least partially. In this study, the high stakes indicators added significantly to the prediction of scale scores, percentage of students reaching Basic, and changes in scale scores over four years (within grade), even after controlling for demographics. Although not significant (due probably to small number of samples in the equation), the high stakes indicators also predicted large proportions of variance (13%) in the change in percentage of students reaching Basic. Interestingly, as with reading achievement, although the presence of high stakes testing was related to lower math achievement, years of testing and the high stakes index were related to higher math achievement in a single year and the high stakes index to gains in math between testing periods. However, none of the high stakes indicators were related to percentage of students reaching Proficient.

We might argue that the degree to which a state has put testing policies in place (as assessed by the high stakes index) and the number of years of testing may be associated with several practices that might explain the results for math proficiency. In particular, Nichols et al. (2006) suggested that high stakes testing policies leads to more drill and practice, a practice that would enhance math but not reading achievement, especially at lower grade levels when math content is more concrete. Similarly, we would assert that states under more accountability pressure most likely have developed curricular standards and have aligned teaching practices with these standards; the use of such standards would be related to the number of years that testing has been in place. Having educational policies that lead to higher achievement (e.g., standards and curricular alignment) is not being disputed. However, these results do not vindicate a general educational reform effort focused almost exclusively on testing nor does it provide adequate support to any argument that high stakes testing is necessary to raise student achievement. In particular, even though some impact may be seen among students at the lower levels (Basic), these results further support the notion that high stakes testing does little if anything for students at higher levels of performance (Fuller et al., 2006). Instead, the results suggest the need to further explore what it is about high-stakes testing policies that might influence mathematics achievement, and at what grade levels.

The purpose of high stakes testing continues to be called into question. In addition to the series of studies by Berliner and colleagues and by Marchant and colleagues, other recent research has failed to support the contention that accountability policies result in a decrease in the achievement gaps related to race and SES (Borman et al., 2004; Lee & Wong, 2004). In light of the expense and unintended negative consequences being identified in the research, the bottom-line question concerning high-stakes testing must be, is high-stakes testing worth it as a general approach to educational reform? If this type of testing did not take much time or much money, the lack of consistent evidence supporting achievement would not be as important. However, the Government Accountability Office estimates that states will spend between $1.9 billion and $5.3 billion in the next six years (Olson, 2004). In a time when states and school districts are facing difficult choices due to financial constraints, every educational expense should be justified. It is time to slow the bandwagon and thoroughly examine what it is about high-stakes testing that is worth the price.

References

American Educational Research Association. (2000). *AERA position statements: High-stakes testing in PreK-12 education.* Retrieved March 27, 2005, from http://www.aera.net/policyandprograms/?id=378.

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

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

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

Borman, K. M., Eitle, T. E., Michael, D., Eitle, D. J., Lee, R., Johnson, L., Cobb-Roberts, D., Dorn, S., & Shircliffe, B. (2004). Accountability in a postdesegregation era: The continuing significance of racial segregation in Florida's schools. *American Educational Research Journal, 41*, 605–631.

Berliner, D. C. (2005). Our impoverished view of educational reform. *Teachers College Record,* August. Retrieved August 23, 2005, from http://www.tcrecord.org ID Number: 12106.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Educational Policy Analysis Archives, 12*(1). Retrieved March 24, 2005, from http://epaa.asu.edu/epaa/v12n1/.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–31.

Corno, L. (2000). Comments on Trojan horse papers. *Issues in Education, 6*, 125–131.

Eccles, J. S. (2005). Influences of parents' education on their children's educational attainments: The role of parent and child perceptions. *London Review of Education, 3*(3), 191–204.

Entwisle, D. R., Alexander, K. L., & Steffel Olson, L. (2005). First grade and educational attainment by age 22: A new story. *American Journal of Sociology, 110*(5), 1458–1502.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the no child left behind act working: The reliability of how states track achievement.* Berkeley, CA: Policy Analysis for California Education, University of California.

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

Jones, L. V., & Olkin, I., (Eds.). (2004). *The nation's report card: Evolution and perspectives.* Bloomington, IN: Phi Delta Kappa.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41,* 797–832.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–15.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, UCLA (CSE Technical Report 539). Retrieved September 6, 2006, from http://www.cse.ucla.edu/Reports/TR539.pdf.

Marchant, G. J. (2005, April). *Aggregation and disaggregation in accountability testing.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, CA.

Marchant, G. J., & Paulson, S. E. (2001). State comparisons of SAT scores: Who's your test taker? *NASSP Bulletin, 85*(627), 62–74.

Marchant, G. J., & Paulson, S. E. (2005). The relationship of high school graduation exams to graduation rates and SAT scores. *Educational Policy Analysis Archives, 13*(6). Retrieved January 21, 2005, from http://epaa.asu/epaa/v13n6/.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Educational Policy Analysis Archives, 14*(1). Retrieved February 3, 2006, from http://epaa.asu/epaa/v14n1/.

Olson, L. (2004, December 1). NCLB law bestows bounty on test industry. *Education Week, 24*(14), 1, 18–19. Retrieved March 21, 2005, from http://www.edweek.org/ew/articles/2004/12/01/14tests.h24.html.

Paris, S. G. (2000). Trojan horse in the schoolyard: The hidden threats in high-stakes testing. *Issues in Education, 6*, 1–16.

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved August 2, 2004, from http://epaa.asu.edu/epaa/v11n24/.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.

Slavin, R. E. (1989). PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan, 70*, 752–58.

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

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461–481.

About the Authors

Gregory J. Marchant, Sharon E. Paulson, and Adam Shunk
Ball State University

Email: gmarchant@bsu.edu

Greg Marchant is a Professor of Psychology—Educational Psychology at Ball State University. His current research focuses on high-stakes testing and the uses and misuses of aggregated test scores. He is also involved in the evaluation and development of the Learning Assessment Model Project used to demonstrate student learning of teacher candidates.

Sharon E. Paulson is a Professor of Psychology—Educational Psychology at Ball State University. She specializes in adolescent development and works closely with secondary education majors on understanding developmental principles important to teaching and learning. Other research interests include the effects of parenting on adolescent achievement.

Adam Shunk is a doctoral student in the School Psychology program at Ball State University. He is currently completing his clinical internship and working on his dissertation.

Appendix A
States' Participation In NAEP Reading, Writing, Math, And Science Testing 1992–2002

Table A1
States' Participation In NAEP Reading And Writing Testing For Each Testing Period 1992–2002

	Reading						Writing		
Year	1992	1994	1998	1998	2002	2002	1998	2002	2002
Grade	4	4	4	8	4	8	8	4	8
Missing Data	$ Edu.	$ Edu.	Edu.		Edu.			Edu.	
Total States	42	40	40	37	44	42	36	44	42
AL	x	x	x	x	x	x	x	x	x
AK	—	—	—	—	—	—	—	—	—
AZ	x	x	x	x	x	x	x	x	x
AR	x	x	x	x	x	x	x	x	x
CA	x	x	x	x	x	x	x	x	x
CO	x	x	x	x	—	—	x	—	—
CT	x	x	x	x	x	x	x	x	x
DC	x	x	x	x	x	x	x	x	x
DE	x	x	x	x	x	x	x	x	x
FL	x	x	x	x	x	x	x	x	x
GA	x	x	x	x	x	x	x	x	x
HI	x	x	x	x	x	x	x	x	x
ID	x	—	—	—	x	x	—	x	x
IL	—	—	—	—	—	—	—	—	—
IN	x	x	—	—	x	x	—	x	x
IA	x	x	x	—	x	—	—	x	—
KS	—	—	x	x	x	x	—	x	x
KY	x	x	x	x	x	x	x	x	x
LA	x	x	x	x	x	x	x	x	x
ME	x	x	x	x	x	x	x	x	x
MD	x	x	x	x	x	x	x	x	x
MA	x	x	x	x	x	x	x	x	x
MI	x	—	x	—	x	x	—	x	x
MN	x	x	x	x	x	—	x	x	—
MS	x	x	x	x	x	x	x	x	x
MO	x	x	x	x	x	x	x	x	x
MT	—	x	x	x	x	x	x	x	x
NE	x	x	—	—	x	x	—	x	x
NV	—	—	x	x	x	x	x	x	x

	Reading						Writing		
Year	1992	1994	1998	1998	2002	2002	1998	2002	2002
Grade	4	4	4	8	4	8	8	4	8
Missing Data	$ Edu.	$ Edu.	Edu.		Edu.			Edu.	
Total States	42	40	40	37	44	42	36	44	42
NH	x	x	x	—	—	—	—	—	—
NJ	x	x	—	—	—	—	—	—	—
NM	x	x	x	x	x	x	x	x	x
NY	x	x	x	x	x	x	x	x	x
NC	x	x	x	x	x	x	x	x	x
ND	x	x	—	—	x	x	—	x	x
OH	x	—	—	—	x	x	—	x	x
OK	x	—	x	x	x	x	x	x	x
OR	—	—	x	x	x	x	x	x	x
PA	x	x	—	—	x	x	—	x	x
RI	x	x	x	x	x	x	x	x	x
SC	x	x	x	x	x	x	x	x	x
SD	—	—	—	—	—	—	—	—	—
TN	x	x	x	x	x	x	x	x	x
TX	x	x	x	x	x	x	x	x	x
UT	x	x	x	x	x	x	x	x	x
VT	—	—	—	—	x	x	—	x	x
VA	x	x	x	x	x	x	x	x	x
WA	—	x	x	x	x	x	x	x	x
WV	x	x	x	x	x	x	x	x	x
WI	x	x	x	x	—	—	x	—	—
WY	x	x	x	x	x	x	x	x	x

States participating in each testing period are indicated with an **x**.

Table A2

States' Participation In NAEP Math And Science Testing For Each Testing Period 1992–2000

	Math						Science		
Year	1992	1992	1996	1996	2000	2000	1996	2000	2000
Grade	4	8	4	8	4	8	8	4	8
Missing Data	$ Edu.	$	Edu.		Edu.			Edu.	
Total States	42	42	44	41	41	40	41	39	38
AL	x	x	x	x	x	x	x	x	x
AK	—	—	x	x	—	—	x	—	—
AZ	x	x	x	x	x	x	x	x	x
AR	x	x	x	x	x	x	x	x	x
CA	x	x	x	x	x	x	x	x	x
CO	x	x	x	x	—	—	x	—	—
CT	x	x	x	x	x	x	x	x	x
DC	x	x	x	x	x	x	x	—	—
DE	x	x	x	x	—	—	x	—	—
FL	x	x	x	x	—	—	x	—	—
GA	x	x	x	x	x	x	x	x	x
HI	x	x	x	x	x	x	x	x	x
ID	x	x	—	—	x	x	—	x	x
IL	—	—	—	—	x	x	—	x	x
IN	x	x	x	x	x	x	x	x	x
IA	x	x	x	x	x	—	x	x	—
KS	—	—	—	—	x	x	—	—	—
KY	x	x	x	x	x	x	x	x	x
LA	x	x	x	x	x	x	x	x	x
ME	x	x	x	x	x	x	x	x	x
MD	x	x	x	x	x	x	x	x	x
MA	x	x	x	x	x	x	x	x	x
MI	x	x	x	x	x	x	x	x	x
MN	x	x	x	x	x	x	x	x	x
MS	x	x	x	x	x	x	x	x	x
MO	x	x	x	x	x	x	x	x	x
MT	—	—	x	x	x	x	x	x	x
NE	x	x	x	x	x	x	x	x	x
NV	—	—	x	—	x	x	—	x	x
NH	x	x	—	—	—	—	—	—	—
NJ	x	x	x	—	—	—	—	—	—
NM	x	x	x	x	x	x	x	x	x
NY	x	x	x	x	x	x	x	x	x
NC	x	x	x	x	x	x	x	x	x

	Math						**Science**		
Year	1992	1992	1996	1996	2000	2000	1996	2000	2000
Grade	4	8	4	8	4	8	8	4	8
Missing Data	$ Edu.	$	Edu.		Edu.			Edu.	
Total States	42	42	44	41	41	40	41	39	38
ND	x	x	x	x	x	x	x	x	x
OH	x	x	—	—	x	x	—	x	x
OK	x	x	—	—	x	x	—	x	x
OR	—	—	x	x	x	x	x	x	x
PA	x	x	x	—	—	—	—	—	—
RI	x	x	x	x	x	x	x	x	x
SC	x	x	x	x	x	x	x	x	x
SD	—	—	—	—	—	—	—	—	—
TN	x	x	x	x	x	x	x	x	x
TX	x	x	x	x	x	x	x	x	x
UT	x	x	x	x	x	x	x	x	x
VT	—	—	x	x	x	x	x	x	x
VA	x	x	x	x	x	x	x	x	x
WA	—	—	x	x	—	—	x	—	—
WV	x	x	x	x	x	x	x	x	x
WI	x	x	x	x	—	—	x	—	—
WY	x	x	x	x	x	x	x	x	x

States participating in each testing period are indicated with an **x**.

EDUCATION POLICY ANALYSIS ARCHIVES http://epaa.asu.edu

Editor: Sherman Dorn, University of South Florida

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Sherman Dorn, epaa-editor@shermandorn.com.