The failure of educational accountability to work as intended in the United States

Sharon L. Nichols
Associate Professor
University of Texas at San Antonio, USA

Abstract

The goal of this paper is two fold. First I review data from studies conducted by colleagues and myself looking at the relationship between high-stakes testing and student achievement. These data suggest that the pressure of high-stakes testing practices impact fourth grade math significantly and eighth grade math to a lesser degree. There is no relationship between high-stakes testing and reading. Second, I review literature looking at the unintended consequences of high-stakes testing. Although there are an increasing number of scholars documenting the mostly deleterious effects of high-stakes testing in teaching and learning, there has been much less focus on how it impacts students—specifically student motivation. I conclude the paper by discussing how important it is that we begin to worry more about the consequences of high-stakes testing when it comes to students and their outlook on learning and future goals.

"Much like medical professionals dealing with cancer, educators are expected to fix problems of teaching and learning by treating symptoms of the problem, though they have only imperfect knowledge and no control of its causes. While schools are part of the solution, they alone cannot solve the problem of educational disparities" (Timar, 2012, p. 230)

Accountability Mechanisms of the United States

In 2001, US President George W. Bush signed into law the No Child Left Behind Act (or otherwise known as NCLB). The goal of NCLB was to improve America's public education system deemed to be failing through a series of federal mandates. At the core of these mandates is that states develop and implement high-stakes testing accountability systems or else lose millions in federal support. These mandates asked states to create statewide standards in all grade levels and subjects and a state wide standardized test to measure student progress against those standards. Lastly, and most importantly states had to implement a set of escalating consequences to be administered when students underperform on these tests. In other words, states had to enforce a system of high-stakes testing whereby teachers and students could earn rewards for scoring well, or be punished for not scoring well. Tests had become the sole indicator of teaching effectiveness.

The theory of action underlying the practice of high-stakes testing suggests that if we pressure teachers and students by threatening them with job losses or grade retention then this external pressure will motivate everyone to do "better." Although several states and some districts have used tests as a lever for school reform and for holding administrators, teachers, and their students "accountable" for quite sometime, it has been a national mandate only since 2002 when NCLB was passed into law. Decades of high-stakes testing implementation have provided scholars an opportunity to examine its effects. The result has been a growing literature base that can be organized by two strands of foci: the intended and unintended impact of high-stakes testing

In terms of the intended impact, research is mixed. There are no consistent data to underscore whether high-stakes testing has had the intended effect of raising student achievement and/or closing the achievement gap. By contrast, in terms of the unintended effects, the literature is growing rapidly to suggest that the practice of high-stakes testing has increasingly eroded and undermined sound educational practice. The goal of this paper is to address both strands data.

Purpose

This paper has two major goals. The first goal is to summarize data from research conducted by my colleagues and myself on the relationship between educational accountability practices and student achievement. If the goal of high-stakes testing is to increase student achievement, then we would expect to see significant positive correlations between an indicator of high-stakes testing and student achievement. In the work I present here, we use a uniquely derived indicator of high-stakes testing practice to examine how those practices relate to student achievement.

A second goal of this paper is to extrapolate from the findings in part one in order to underscore the range of unintended effects of high-stakes testing. There is rapidly growing data to suggest that as the pressure to perform on test rises, so too do undesirable educational practices. Others and myself have commented the nature and type of these practices on widely. My goal here is to extend upon these discussions to comment on how these practices impact students. Extant data too often focus on instruction and other school and/or classroom-level concerns. I would like to add to these discussions a concern for the student and not only their academic outcomes, but also their motivational outcomes.

Review of Relevant Literature

High-stakes testing is the process of attaching significant consequences to standardized test performance with the goal of incentivizing teacher effectiveness and student achievement (Herman & Haertel, 2005; Ryan, 2004). The rationale is that by attaching significant rewards or serious threats to changes in student test scores, teachers and their students will inevitably be prompted to work harder, better, and learn more. Although most tests students take are arguably "high-stakes" to them (i.e., failing a teacher-made test could result in failing a class or not passing to the next grade),"high-stakes" here refers to standardized tests developed specifically for the purpose of evaluating teachers and students. Performance on these tests may result in important consequences to schools, administrators, teachers, and students. Passing could bring rewards to teachers (bonuses) and schools (positive reviews in local newspapers), whereas failure could bring severe penalties to teachers and principals (termination), schools (closure or "take-over"), and to students (denied diploma or retained in grade).

Although the practice of high-stakes testing gained a prominent position in educational reform with the passage of the No Child Left Behind Act (NCLB) of 2002, its use as a lever for school change preceded NCLB. Tests have been used to distribute rewards and sanctions to teachers in urban schools since the mid 1800s (Tyack, 1974) and for most schools throughout the United States since at least the 1970s (Haertel & Herman, 2005). New York state in particular has led the United States in test-based accountability efforts, "implementing state-developed (1965) and mandated minimal competency testing (MCT) before most other states (1978) and disseminating information to the media about local district performance on the state assessments before it became routinely popular (1985)" (Allington & McGill-Franzen, 1992, p. 398).

Standardized achievement test invention, development, and use paralleled these reform efforts (Giordano, 2005). The evolution of valid and reliable measurement techniques influenced views of how one might gauge educational quality (McDonnell, 2005). The passage of NCLB in

2002, mandated the most intrusive use of tests for influencing how and what teachers would teach and how and what students would learn. In spite of a growing literature indicating that high-stakes testing has had deleterious effects on teaching practices and student motivation, policymakers continue to argue for its effectiveness in increasing student learning as evidenced in newer proposals (e.g., U.S. Department of Education, 2009) and recommendations for the reauthorization of NCLB (U.S. Department of Education, 2010).

High-Stakes Testing and Student Achievement

Most of the research conducted around the time of NCLB provides scant support for the effectiveness of high-stakes tests in increasing student achievement (Amrein & Berliner, 2002a, b; Braun, 2004; Rosenshine, 2003) or graduation rates (Haney *et al*., 2004; Heubert & Hauser, 1999; Marchant & Paulson, 2005). And, since our first study on this topic (Nichols, Glass & Berliner, 2006), no data have emerged to contradict the findings that accountability pressure has some relationship to fourth grade math, virtually no influence on reading (Dee & Jacob, 2009), and only negative influence on student graduation rates (Holme, Richards, Jimerson, & Cohen, 2010; Orfield, Losen, Wald, & Swanson, 2004). Studies focusing on both high- and low-stakes exit exams repeatedly reveal that these types of incentive/threat have little to no impact on student achievement over time (e.g., Bishop, Mane, Bishop, & Moriarty, 2001; Grodsky, Warren, & Kalogrides, 2009; Reardon, Arshan, Atteberry, & Kurlaender, 2008; Reardon, Atteberry, Arshan, & Kurlaender, 2009). In addition, the reduction of the achievement gap between income groups and between racial and ethnic groups, a major goal of the high-stakes accountability movement, either did not occur or was only marginally effective in the years these policies have been in place (Braun et al., 2006, 2012; Reardon, 2011; Timar & Maxwell-Jolly, 2012).

Unintended Effects of High-Stakes Testing

Although it is difficult to conclude with certainty the exact nature of the relationship between pressure and student achievement, it is exceedingly easier to make conclusions regarding the ways in which high-stakes testing has impacted teaching and learning practices. In the years since NCLB and other high-stakes testing reform efforts were already underway, scholars have documented (both quantitatively and qualitative) how accountability-based testing has impacted actual classroom- and/or school-wide practices. In our own work, colleague David Berliner and I uncovered a wide range of corrupting practices as the result of high-stakes testing pressure. Using what we referred to as Campbell's law, we carefully documented these practices in our book *Collateral Damage* (Nichols & Berliner, 2007).

Campbell's law, named after the well-respected social psychologist, evaluator, methodologist, and philosopher of science Donald Campbell, brought his adage to the attention of social scientists decades ago in a 1975 edited book about evaluation. Campbell's law stipulates that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it was intended to monitor." (Campbell, 1975, p. 36) Campbell warned us a long time ago of the inevitable problems associated with undue weight and emphasis on a single indicator for monitoring complex social phenomena. In effect, he warned us about the corruption that would ensure under the weigh of our current high-stakes testing program that is part and parcel of NCLB.

George Madaus and Margarite Clarke furthered Campbell's warning arguing (Madaus & Clarke, 2001) that Campbell had uncovered a kind of social science equivalent of the Heisenberg uncertainty principle. Named after its discoverer, Werner Heisenberg, the Heisenberg Uncertainty Principle states that there is always a degree of uncertainty associated with simultaneously measuring the position and velocity of microscopic objects. Heisenberg explains,

"The more precisely the position is determined, the less precisely the momentum is known in this instant, and vice versa." (Heisenberg, 1927). To try to measure a microscopic object inevitably alters the conditions so much as to render the measurement inaccurate—there is inherent uncertainty in the data that is collected.

Madaus and Clarke noted that if Campbell is right, as he seems to be, whenever you have high-stakes attached to an indicator, such as test scores, you have a measurement system that has been corrupted, rendering the measurement less accurate. You apparently can have higher stakes and less certainty about the validity of the assessment, or lower stakes and greater certainty about validity. But you are not likely to have both high stakes and high validity. Uncertainty about the meaning of test scores increases as the stakes attached to them become more severe. The higher the stakes the more likely it is that the construct being measured has been changed. Applied to NCLB, scores on the instruments used as measures of academic achievement become less interpretable as the consequences for not making adequate yearly progress increase.

Two Studies on the Intended Impact of High Stakes Testing: Methods Overview

Study One

Our first study on the relationship between high-stakes tests and student achievement (Nichols et al., 2006) was prompted by our view that existing approaches to the *measurement* of test-based accountability policies and practice at the state level were largely inadequate because of their reliance on inspection of state level legislation, as opposed to actual practices (e.g., Carnoy & Loeb, 2002; Rosenshine, 2003; Swanson & Stevenson, 2002). That is, most researchers measured testing "pressure" by examining the number of laws that states had passed prior to or up to the enactment of NCLB. Although reliable, the validity of such approaches for capturing the on-the-ground feeling of pressure was doubtful. In this first study, we addressed

this problem by spending considerable time and effort conceptualizing and building a measure that would more closely represent high-stakes testing policy implementation and which we labeled the Accountability Pressure Rating (APR).

Our method for deriving APR values for our 25[1] study states was guided by the method of "comparative judgments" used for ordering complex and abstract psychological data (Torgerson, 1960). This approach seemed ideal for our purposes since our goal was to transform complex qualitative data (state level policy legislation enactment and implementation) into a quantitative indicator that can be used in subsequent analyses. Our work involved three steps. First, we created state-level portfolios that included a range of legislative documentation, state-generated accountability reports, and newspaper articles documenting the range of ways policy changes both impacted and were viewed by the public (see Nichols et al., 2006 for a complete description of portfolio contents). One unique aspect of our approach was the inclusion of newspaper references (both leading stories as well as editorials) that were used to capture the on-the-ground effects of and reactions to local and statewide test-based accountability practices. In contrast to other studies that relied on quantitative estimations of policies (e.g., Braun, 2004; Carnoy & Loeb, 2002) our measure included evidence that described how policies played out in local school systems.

Next, we asked 300 graduate students each to view two states' portfolios and to make two judgments—which state exerted more pressure and by about how much (on a scale of 1-7[2]).

---

[1] 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—the 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.

[2] We felt a 7-point scale would provide enough variability in raters' responses while at the same time minimize reliability issues emanating from wider scales.

Last, we took the ratings provided by our students and applied the least-squares solution for uni-dimensional scale values due to Mosteller (as outlined in Torgerson, 1960, pp. 170–173). The result was a scale ranging from .54 to 4.78. This rating served as our measure of state-level testing pressure as of 2004 (See Table 1). As can be seen in Table 1, Kentucky's policies and practices were consistently rated below other states in terms of test-based pressure (APR = .54), whereas Texas's policies and practices were consistently viewed as having the highest test-based accountability pressure (APR= 4.78). Using our APR, we performed correlation and regression analyses to examine patterns in the relationships between our APR and fourth and eighth grade reading and math NAEP[3] through 2003.

Table 1: Accountability Pressure Rating (APR, 2004)

State	APR	State	APR	State	APR
Kentucky	0.54	Utah	2.80	Georgia	3.44
Wyoming	1.00	Maryland	2.82	Tennessee	3.50
Connecticut	1.60	Alabama	3.06	Louisiana	3.72
Hawaii	1.76	Virginia	3.08	Mississippi	3.82
Maine	1.78	West Virginia	3.08	New York	4.08
Rhode Island	1.90	Massachusetts	3.18	NC	4.14
Missouri	2.14	SC	3.20	Texas	4.78
California	2.56	New Mexico	3.28		
Arkansas	2.60	Arizona	3.36		

Study Two

In our follow up study, we looked at the relationship between our APR and later NAEP data. Specifically we wanted to know, what is the pattern of correlations between APR and fourth and eighth grade NAEP scores in reading and math from 2005-2009:

[3] The National Assessment of Educational Progress (NAEP) is the largest nationally representative and continuing assessment of what America's students know and can do in various subject areas. Assessments are conducted periodically in mathematics, reading, science, writing, the arts, civics, economics, geography, and U.S. history.

- over time;
- when disaggregated by student ethnicity;
- when disaggregated by student socioeconomic status;

Additionally, what is the relationship between APR and four-year NAEP gains (both cohort and non-cohort) in math and reading?

In both studies, we used descriptive statistics to analyze fourth and eighth grade NAEP data during the relevant study periods in reading and math. We conducted a series of partial, part, and simple bivariate correlations to examine relationships among state level demographic characteristics, APR, and NAEP indicators.

Selected Results

From both studies, we found that our APR was connected most consistently with gains in fourth grade math performance, only slightly connected to gains in eighth grade math, and not correlated with gains in reading at either fourth or eighth grade levels (Nichols et al., 2006, Nichols et al., 2012). In combination, our data yielded total of 51 tables. I share two of these here.

In both studies, we concluded with a table that rank ordered all correlations (bivaraite, and partial) emanating from analyses in which we looked at how our APR connected to student NAEP achievement disaggregated by ethnicity. [4] The result was a pattern of correlations that yielded an interesting pattern. In Table 2 I display this rank ordering from our 2006 study and in Table 3 from our 2012 study. Data displayed in Table 2 represents analyses we did where we calculated the hypothesized change in test-based pressure in years preceding NAEP gain scores. These correlations include analyses of preceding pressure changes (1988-1992; 1992-1996, and 1996-2000) as they related to subsequent non-cohort NAEP gain scores (1992-1996; 1996-2000, and 2000-2003). Our logic was that although our data were correlation in nature, this antecedent-

[4] For specific information regarding the formulas used, please consult original studies.

consequent correlation design would provide some clues regarding a causal relationship. As can be seen in Table 2, our data yielded 40 correlations that are rank ordered according to absolute value. By dissecting these 40 correlations in half, among the weakest/negative correlations are 8 math relationships and among the stronger, positive correlations are 13 math relationships. The strongest, positive correlations all come from fourth grade math and the strongest negative correlations come from fourth grade reading.

In Table 3, I display correlations rank ordered by absolute values emerging from analyses where student achievement was disaggregated by ethnicity. These 48 correlations reveal a similar pattern as is shown in Table 2. Among the bottom (more negative) 24 correlations, 19 come from reading and 5 from math. By contrast, among the top (more positive) 24 correlations 19 come from math and 5 from reading. Positive relationships between pressure and NAEP performance exist primarily in math across all subgroups. In contrast to data from Table 2, in which we found significant relationships between APR and NAEP in fourth grade math primarily, these correlations seem more evenly spread among fourth and eighth grade performance.

Implications From the Data

The research on the impact of accountability-based policies and student achievement is varied, limited, and relatively inconclusive. One explanation for this state of affairs is that it is very difficult to isolate cause-effect relationships between complex policy implementation and subsequent student achievement. Still, our data here and elsewhere, as well as work by others reiterate a familiar story: Increased testing pressure is related to increases in achievement in math more consistently than in reading. Differences in the nature of the mathematics and reading curriculum, and /or differences in the ways one can prepare for assessments in these two areas may have something to do with the fact that state level pressure to perform well on high-stakes

tests is more strongly and positively related to math achievement and negatively related to reading achievement.

Table 2: Antecedent—Consequent Change Correlations for Various Subjects, Ethnicities and Grades: Non-Cohort Analyses (1995-2003)

-0.38	**4th grade reading, AA**
-0.3	**4th grade reading, Hispanic**
-0.18	8th grade math, White
-0.16	4th grade math, White
-0.11	8th grade math, AA
-0.11	4th grade reading, Hispanic
-0.09	4th grade math, AA
-0.05	4th grade reading, White
-0.02	8th grade math, AA
0	8th grade reading, AA
0	8th grade math, AA
0	4th grade math, White
0.01	4th grade reading, AA
0.01	8th grade reading, White
0.04	8th grade reading, AA
0.04	4th grade reading, White
0.06	4th grade reading, Hispanic
0.08	8th grade math, Hispanic
0.08	8th grade reading, White
0.08	4th grade reading, AA
0.12	4th grade reading, Hispanic
0.12	4th grade reading, White
0.12	8th grade reading, White
0.13	4th grade reading, AA
0.15	8th grade math, White
0.15	8th grade reading, Hispanic
0.16	8th grade math, AA
0.16	8th grade math, Hispanic
0.16	8th grade reading, Hispanic
0.18	4th grade math, Hispanic
0.25	4th grade math, Hispanic
0.25	8th grade math, Hispanic
0.28	4th grade reading, White
0.3	8th grade math, White
0.31	8th grade math, Hispanic
0.33	8th grade math, White
0.37	**4th grade math, AA**
0.42	**4th grade math, Hispanic**
0.43	**4th grade math, AA**
0.73	**4th grade math, White**

Table 3: Rank ordering of APR-NAEP correlations disaggregated by student ethnicity

Student Subgroup	APR/NAEP Correlation	Grade	Subject	Year
Hispanic	.463	8	Math	2005
African American	.410	8	Math	2007
Hispanic	.399	4	Math	2007
African American	.390	4	Math	2005
Hispanic	.362	4	Math	2005
African American	.358	8	Math	2005
African American	.324	8	Reading	2009
Hispanic	.312	8	Math	2007
White	.308	4	Math	2003
White	.297	4	Math	2005
White	.294	8	Math	2005
Hispanic	.291	4	Math	2003
African American	.290	8	Math	2003
African American	.282	4	Math	2003
Hispanic	.259	8	Math	2003
White	.254	8	Math	2003
African American	.242	8	Reading	2007
White	.240	4	Math	2007
White	.225	8	Math	2007
Hispanic	.217	8	Reading	2009
White	.211	8	Math	2009
African American	.191	4	Math	2007
Hispanic	.190	4	Reading	2007
Hispanic	.170	8	Reading	2003
Hispanic	.162	4	Math	2009
African American	.152	8	Math	2009
White	.149	8	Reading	2007
White	.143	8	Reading	2003
White	.140	4	Reading	2003
White	.120	4	Math	2009
Hispanic	.098	4	Reading	2005
White	.098	4	Reading	2005
African American	.091	4	Reading	2009
Hispanic	.063	8	Math	2009
African American	.038	4	Math	2009
White	.030	8	Reading	2005
White	.005	4	Reading	2007
African American	.004	8	Reading	2005
White	.003	4	Reading	2009
African American	-.017	4	Reading	2007
White	-.021	8	Reading	2009
African American	-.028	4	Reading	2005
African American	-.059	8	Reading	2003
Hispanic	-.075	4	Reading	2003
Hispanic	-.077	8	Reading	2007
Hispanic	-.120	8	Reading	2005
African American	-.169	4	Reading	2003
Hispanic	-.177	4	Reading	2009

Limitations of the Data

Importantly, there are limitations to what our data offer. In both studies, analyses are correlation in nature and therefore, it is difficult to pinpoint the causal nature of the relationship. Although in our first study, we used an antecedent-consequent analytic strategy that allows us to closer approximate causality we still must interpret these causal relationships with caution. In our second study, a significant limitation has to do with our APR measure. APR was derived from 2004 data and in our second study we wanted to understand how pressure related to later NAEP achievement (2005-2009). We cannot say with certainty whether states' relative APR position against other states holds up over time. Although there have been few changes in the high-stakes testing laws during the 2005-2009 time frame that would fundamentally alter any given state's high-stakes testing positioning, the major question is whether states changed relative to one another over time.

Moving Forward: Thinking About Students

From our data, we conclude that the overall pattern of correlations (math more strongly connected to pressure than reading), points to the likelihood that under pressure, teachers grow more efficient at training students for the test. Although our data cannot conclusively prove this point, the growing amount of anecdotal data regarding how teaching has changed in the face of such pressure provides greater support (e.g., Nichols & Berliner, 2007; Perlstein, 2007). Therefore, it is important that we begin to more critically examine how the changing educational climate under the weight of high-stakes testing is affecting our youth. As a motivational researcher, I extremely worried about how the testing culture is socializing youth and their developing motivational dispositions towards school and learning.

High-Stakes Testing Impacts Contexts in Which Students Learn

There are growing data to demonstrate that the institution of high-stakes testing throughout the American public school system has caused an indelible shift in school/classroom contexts. In virtually every public school, administrators and their teachers have become increasingly worried about how students will perform on these mandated high-stakes tests. As a result, the way in which teachers teach, structure the daily curriculum and relate to their students throughout each academic year is almost wholly dictated by how well students perform on the test (Au, 2007). In short, annual high-stakes testing has changed what gets taught, the messages communicated throughout a school about what it means to be a member of the school community, and how teachers relate to their students.

What gets tested is what gets taught

Increasingly, and as a result of the pressures associated with high-stakes tests we see the curriculum shifting to reflect what will be on the test. As a result, the curriculum to which students are exposed has become narrower, dryer and less interesting. In terms of a narrowed curriculum, data reveal that subjects not on the test are increasingly cut from the curriculum, while subjects on the test receive more time throughout a school day and year (e.g., Zastrow & Janc, 2006). Surveys conducted by the Center on Education Policy (CEP) with a nationally representative sample of school districts found that since from 2000-2001 to 2006-2007 (and the federally mandated institution of high-stakes testing), 62% of them had increased the amount of time dedicated to English Language Arts (ELA) and Math (most likely to be tested) in elementary schools. And, on average, districts reported an increase of 141 and 89 weekly minutes added to instructional time for ELA and math (respectively) (McMurrer, 2008; see also Nichols & Berliner, 2007).

What is especially problematic is that these curriculum-narrowing activities to improve test scores also result in a "dummying down" effect. Teachers, pressured to prepare students to

perform in math and reading, engage in repetitious instruction that boils down content to isolated bits of information, leaving little time left to engage in creative interdisciplinary activities or project-based inquiry. One Colorado teacher put it this way,

> Our district told us to focus on reading, writing, and mathematics….in the past I had hatched out baby chicks in the classroom as part of the science unit. I don't have time to do that. I have dissected body parts and I don't have time to do that….we don't do community outreach like we used to, like visiting the nursing home or cleaning up the park that we had adopted." (Taylor, Shepard, Kinner, & Rosenthal, 2003, p. 51).

Another teacher offered this perspective, "I'm teaching more test-taking skills and how to use your time wisely. Also what to look for in a piece of literature and how to underline important details. There is a lot more time spent on teaching those kinds of skills…Read questions, restate the question in your answer, how to write so the person grading the test can read it, etc." (Taylor, et al., 2003, p. 39).

Exaggerated importance of tests

Tests are also made too important throughout school cultures. For example, anecdotal data reveal that tests are made important through pep rallies, ice cream socials, and other peculiar events meant to "motivate" students to do well on the test. For example, one Texas high school held a rally for parents, teachers, and students during which the principal informed parents of the importance of the Texas high-stakes test (TAKS) and how it was like a marathon, in which "students need endurance." He was not subtle when he told parents, "This is the test of your lives!" (Foster, 2006). Bulletin boards, posters, and daily mantras constitute another form of explicit emphasis on the test's importance. Schools also put cliché slogans on posters and banners throughout the school. Messages like "Take us to Exemplary" or "EAT the TAKS" are pervasive in many Texas schools. Under high-stakes testing, the test becomes the most

important reason for doing well in school, and administrators and their teachers strategically (and persistently) communicate this to students (Nichols & Berliner, 2007, see also Perlstein, 2007).

Students as test scores

The test's importance is also conveyed implicitly through the activities of teachers as they relate to the youth they instruct. In my work, I have found that high-stakes testing causes students to be seen not in terms of their potential or what they bring that is unique or new to the learning environment. Instead, they are too often seen as test score increasers or test score suppressors (Nichols & Berliner, 2007). Students quickly pick this up and realize their test scores define them. Schools identify winners and losers, the privileged and the damned, and they do so on the basis of test scores. Seen as test increasers, students often get "used", such as when they are asked to take exams they have already passed again in order to raise their school's average, or they are pressured to take the test even when they are sick or unavailable. The emphasis on the test leads to cynical attitudes about the point of being in school at all. As one student points out, "The TAKS is a big joke. No really—when no one's looking, we students all laugh. Yep that's the truth…..*this is the easiest test we could ever take*...I mean forget logarithms and algebra, forget knowing about government and what's listed in the Bill of Rights. Instead, we read a two-page story and then answer 11 short questions about it such as, "What was the meaning of the word 'futile' in paragraph two? A: generous, B: deceptive, C: useless and D: applesauce." (High School Junior, quoted in *San Antonio Express-News*, 2007)

For the score suppressors, teachers and principals have been found to do all sorts of unprofessional things to ensure the suppressor either passes (because of rigorous test prep activities, though sometimes by even more questionable means), or is dropped from testing all together. Birmingham, Alabama was an especially notable case in which more than 500 low scoring students were administratively "dropped" from school just days before state testing.

Scores in Birmingham did rise, principals received substantial bonuses for the increased scores, and over 500 students had their lives made infinitely more difficult in the process (Orel, 2003). Booher-Jennings, (2005) found that testing pressures caused teachers to focus only on their "bubble" students (those who were close to passing the test but had not yet), which directed attention away from their highest and lowest achieving students.

A clear message is being sent out all over the country to students: they are valued (or not) for what their test scores can do for the school. High scorers are taken for granted, low scorers are not as valued as high scorers, and bubble students are caught in the middle. The impact of these changing contextual tides is not yet understood when it comes to students' cognitive, affective, or social motivation. This study is designed to examine how these *social processes* unfold in school and classroom contexts.

Students Weigh In on High-Stakes Testing

In spite of the wealth of data emerging on the ways in which high-stakes testing changes educational contexts, there is a paucity of research looking at students' affective, cognitive, and social responses to these changing climates. Still, a review of the available data informs my hypothesis that high-stakes testing erodes (or enhances) student motivation through the ways in which teachers' respond to testing pressure. When many students see education as punitive, uninteresting, and have their abilities narrowly defined by a single test score, the potential for irreparable and damaging consequences is high. For students struggling because education is already a significant challenge, high-stakes testing likely diminishes their sense of self worth and leads to decreased motivation to do well in school, while students who see the high-stakes test as an easy rite of passage view a school culture formed around high-stakes testing as boring and unconnected. A selection of students' voices from Texas confirms these assertions.

Students from a large urban district in Texas understand that adults are trying to sell them the test as the most important reason for learning and for being in school, and it frustrates them. "Students (teachers as well) focus on only the TAKS. It's almost as if they have been given an ultimatum: either pass the test and get the ticket out of there, or pass the test months later and live with the disappointment all your life. It's not fair" (High School Senior, quoted in *San Antonio Express-News*, 2007). Others found the tests dehumanizing and felt frustrated about the narrow curriculum being forced upon them: "The thing I simply despise about this test is how it limits students from expressing themselves. Many times this year I've written papers I know were outstanding, but didn't get the grade I deserved because they didn't meet the 'TAKS format.' On other assignments I'll get the right answers, but won't get credit because it didn't live up to the expectations of the TAKS test…the TAKS test is like plague. Instead of making learning fun and easy, it's slowly killing our education." (High School Freshman, quoted in *San Antonio Express-News*, 2007)

Tests and Student Motivation: What Do We Know?

There is a wealth of empirical and theoretical work in the field of motivation that strongly would suggest that high-stakes testing practices would erode student motivation. At the most simplest level of analysis, we know that students care more about what they learn and are more motivated when they are driven from intrinsic interests and goals. When students care and are interested, they tend to persist and work harder in the face of difficulty. By contrast, when students are immersed in extrinsic climates, it has the potential to erode intrinsic values and diminish effort and persist. On a very basic level, high-stakes testing cultures pervade classrooms and schools with an extrinsic value system—telling students that what is important is the test at all costs. This is costly for high-achieving students who tend to give up in their classes after they know they can pass the test easily. By contrast, it is costly for low achieving students who see the

test as an impossible barrier and then who give up completely. For those students in the middle, it is difficult to predict their reactions, but literature from studies in motivation provides some clues.

Tests inform students what to value

Studies have shown that the form and content of tests exerts a powerful influence on what is important to learn (i.e., informing students what to *value*). According to Rogers (1969, as stated in Crooks, 1988, page 445),

> Examinations tell them [students] our real aims, at least so they believe. If we stress clear understanding and aim at a growing knowledge of physics, we may completely sabotage our teaching by a final examination that asks for numbers to be put into memorized formulas. However loud our sermons, however intriguing the experiments, students will judge by that examination—and so will next year's students who hear about it.

Research also suggests that test content may influence students' study habits and effort (Natriello & Dornbusch 1984; Snyder, 1971). For example, Snyder (1971) noticed that although classroom curricula may emphasize meaning, depth, and problem solving, if the test emphasized rote memorization, then students, wanting to perform well on tests, would often disregard classroom activities focusing on problem-oriented learning and focus on rote memorization to optimize time, effort, and academic success on tests (see also Fredericksen, 1984).

Tests are important mechanisms that convey to students not only *what* they should know, but also *how* they should know it (Bloom, 1956; Krathwohl, 2002). Research suggests, for example, that students vary in their capacity to recognize the level of processing demands made by a test as well as their ability to adapt to those demands. Miller and Parlett (1974) found that some students were highly adept "cue seekers;" those who actively noticed the features of test questions and adapted their study habits accordingly to maximize their test performance. By

contrast, others were "cue conscious," or those who were less active in seeking out test features but still relatively conscious about test-related cues that were handed to them (i.e., by the teacher). Others have shown that students who generally use "surface level" processing approaches (rote memorization) to tests have difficulty adjusting when deeper-level processing is necessary (Martin & Ramsden, 1987). But perhaps more worrisome are data that suggest students capable of deeper level processing switch to surface level approaches when faced with surface-level learning environments (i.e., in classrooms that emphasize rote memorization) (Crooks & Mahalski, 1986).

Elsewhere, Entwistle and Kozeski (1985) examined the question of how curriculum content and evaluation practices set forth through national policy prescriptions impact student study habits in two different countries: Britain and Hungary. At the time, Britain's educational culture was steeped in a summative standardized testing system that emphasized "correctness" whereas Hungarian culture had prioritized creativity and higher order processing in their schools. Using surveys to measure students' study strategies with approximately 1200 13-17-year-olds in Hungary (n=579) and Britain (n=614), Entwistle and Kozeski (1985) found main effect differences in how students approached studying. Britain's students were more apt to employ surface level strategies in learning (memorization) while Hungarian students were more likely to emphasize deeper learning strategies. Entwistle and Kozeski (1985) cautiously conclude that educational assessment environments may influence students' approach to learning.

From this data, it seems as if tests' inherent utility value as a doorway to academic advancement and personal satisfaction (e.g., doing well leads to better grades, pleases the teacher) connects to students' effort and study strategies. Students seem to adapt their study strategies according to the nature and content of the test they face and want to pass. Of course, students do not respond in a monolithic way. Some students engage in adaptive, deep processing

study strategies even if the test promotes memorization and rote learning (Crooks & Mahalski, 1986). Still, testing content conveys important message to students about what the culture values and many adjust their efforts accordingly.

A Final Comment on Student Motivation

According to Expectancy X Value theory (EV) of motivation, students are more likely to persist and to care about their learning when they have expectations for success and value for what they learn. In a high-stakes testing context, these dimensions of motivation are significantly undermined. When students repeatedly fail difficult tasks or repeatedly succeed on easy ones, than their meaningful expectations of success are altered. Similarly, when the value of what students learn is over exaggerated or handed to them without giving students the opportunity to discover their own personal values and interests in a learning environment, then motivation is eroded. It is important that scholars and politicians consider the costs to students under the weight of policies that have such undesirable effects on teaching practice.

Conclusion

Currently, we have a wealth of data regarding the effects of high-stakes testing. When it comes to achievement, results are mixed. Some studies have found significant connections between high-stakes testing practice and student achievement, whereas others have found weak connections. Common to most research, in spite of differences in methods, is the relationship between high-stakes testing and math achievement. Pressure and math performance are positively related, especially at the fourth grade level. By contrast, pressure is not related to reading achievement at any level in any consistent manner. This pattern of results, however, raises some questions. It seems relatively clear that the pressure to perform on tests has some kind of effect on how fourth grade math curriculum is taught. The question remains, however, is what those practices look like. Are teachers becoming ‘better’ math instructors? Or are teachers

and their students becoming better trained at taking fourth grade math tests? Both answers are equally plausible. The growing number of reports from teachers suggests that under pressure, they are narrowing their curriculum and teaching to the test much more frequently. This strongly suggests that the more likely effect of test-based pressures is that teachers and students are becoming better test takers. This of course, is very problematic as under these questions, it is unclear what resultant increases in achievement test scores actually represent.

When it comes to the unintended outcomes of teaching and learning, there is strong evidence to suggest the practice of education is being corrupted in ways that erode or at least undermine the quality of students' experiences. Lesser known are the effects of these negative outcomes on students. Specifically, we really do not know how a culture of test-based pressure is affecting students' long term beliefs about themselves or their futures. The brief review of data here suggests that if teachers overemphasize a test's importance, it will have undesirable effects on students' orientations towards studying and learning. More research is needed to underscore how test-based pressures impact students.

References

Allington, R., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy*, *6*(4), 397–414.

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 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 from http://epaa.asu.edu/epaa/v10n18/

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Bishop, J. H., Mane, F., Bishop, M., & Moriarty, J. (2001). *The role of end-of-course exams and minimum competency exams in standards-based reforms. Brookings Papers on Educational Policy, 4*, 267-345.

Bloom, B. S. (Ed.) (1956). *A taxonomy of educational objectives: Handbook I, the cognitive domain*. NY: Longman.

Booher-Jennings, J. (2005). Below the bubble: 'Educational triage' and the Texas accountability system. *American Educational Research Journal, 42*(2),

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

Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006) The Black-White achievement gap: Do state policies matter? *Education Policy Analysis Archives, 14(8)*. Retrieved from http://epaa.asu.edu/epaa/v14n8/.

Braun, H., Chapman, L., & Vezzu, S. (2010). The Black-White achievement gap revisited. *Education Policy Analysis Archives,18*(21). Retrieved from http://epaa.asu.edu/ojs/article/view/772

Campbell, Donald, "Assessing the Impact of Planned Social Change," in *Social Research and Public Policies: The Dartmouth/OECD Conference*, ed. Gene Lyons (Hanover, NH: Public Affairs Center, Dartmouth College, 1975).

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

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438-481.

Crooks, T. J., & Mahalski, P. A. (1986). Relationships among assessment practices, study methods, and grades obtained. In J. Jones & M. Horsburgh (Eds.), *Research and development in higher education: Vol. 8.* Sydney, Australia: Higher Education Research and Development Society of Australasia.

Dee, T., & Jacob, B. (2009, November). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Entwistle, N.J., & Kozeski, B. (1985). Relationship between school motivation, approaches to studying, and attainment, among British and Hungarian adolescents. *British Journal of Educational Psychology, 55*, 124-137.

Foster, S. L. (2006, December). *How Latino students negotiate the demands of high-stakes testing: A case study of one school in Texas*. Unpublished doctoral dissertation, Arizona State University.

Fredericksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessmen*t. New York: Peter Lang.

Grodsky, E. S., Warren, J. R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971-2004. *Educational Policy, 23*, 589-614. doi: 10.1177/0395909808320678

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In. J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement*. The 104th Yearbook of the National Society for the Study of Education (part 2, pp. 1-34). Malden, MA: Blackwell. doi: 10.1111/j.1744-7984.2005.00023.x

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.

Heisenberg, Werner, quoted in David Cassidy, "Quantum Mechanics 1925-1927: Implications of Uncertainty," *American Institute of Physics Web Site*, http://www.aip.org/history/heisenberg/p08.htm.

Herman, J. L., & Haertel, E. H. (Eds.) (2005). *Uses and misuses of data for educational accountability and improvement*. The 104th Yearbook of the National Society for the Study of Education (part 2). Malden, MA: Blackwell.

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

Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research*, *80*(4), 476-526. doi:10.3102/0034654310383147

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice, 41*(4), 212-218

Madaus, George & Clarke, Marguerite, "The Adverse Impact of High-Stakes Testing on Minority Students: Evidence from One Hundred Years of Test Data," in *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*, eds. Gary Orfield and Mindy L. Kornhaber (New York, NY: Century Foundation Press, 2001).

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

Martin, E., & Ramsden, P. (1987). Learning skills and skill in learning. In J. T. E. Richardson, M. W. Eysenck, & D. W. Piper (Eds.), *Student learning: Research in education and cognitive psychology*. Milton Keynes, England: Open University Press & Society for Research into Higher Education.

McDonnell, L. (2005). Assessment and accountability from the policymaker's perspective. In J. L. Herman, & E. H. Haertel (Eds.). *Uses and misuses of data for educational accountability and improvement: The 104th yearbook of the national Society for the Study of Education, Part II.* Malden, MA: Blackwell

McMurrer (2008). *Instructional time in elementary schools: A closer look at changes for specific subjects*. Washington, DC: Center on Education Policy.
Miller, C. M. L., & Parlett, M. (1974). *Up to the mark: A study of the examination game.* London: Society for Research into Higher Education.
Natriello, G., & Dornbusch, S. M. (1984). *Teacher evaluative standards and student effort.* NY: Longman.
Nichols, S., & Berliner. D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge, MA: Harvard Education Press.
Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved July 20, 2009, from http://epaa.asu.edu/epaa/v14n1/.
Nichols, S. L., Glass, G. V, & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analysis with NAEP data. Education Policy Analysis Archives, 20 (20), retrieved from http://epaa.asu.edu/ojs/article/view/1048
Orel, S. (2003). Left behind in Birmingham: 522 pushed-out students. In R. Cossett-Lent and G. Pipkin (Eds.), *Silent no more: Voices of courage in American schools*. Portsmouth, NJ: Heinemann.
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.
Perlstein, L. (2007). *Tested: One American school struggles to make the grade.* NY: Henry Holt & Co.
Reardon, S. F. (2011). The Widening Academic Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations In Richard Murnane & Greg Duncan (Eds.), Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children. New York: Russell Sage Foundation.
Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2008). *High stakes, no effects: Effects of failing the California High School Exit Exam* (Working Paper 2008-10). Stanford, CA: Stanford University, Institute for Research on Education Policy & Practice.
Reardon, S. F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009, April 21). *Effects of the California High School Exit Exam on Student Persistence, Achievement and Graduation* (Working Paper 2009-12). Stanford, CA: Stanford University, Institute for Research on Education Policy & Practice.
Rosenshine, B. (2003, August 4). High-Stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved from http://epaa.asu.edu/epaa/v11n24/
Ryan, J. E. (2004). The perverse incentives of the No Child Left Behind Act. *New York University Law Review, 79*, 932-989. doi: 10.2139/ssrn.476463
San Antonio Express-News (2007, March 9). Teen Talk: Tackling TAKS: *San Antonio Express-News,* p. F1, 5.
Snyder, B. R. (1971). *The hidden curriculum*. Cambridge, MA: M.I.T. Press
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.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. (CSE Technical Report 588: CRESST/CREDE/University of Colorado at Boulder). Los Angeles: University of California.

Timar, T. B. and Maxwell-Jolly, J. (Eds) (2012). *Narrowing the achievement gap: Perspectives and strategies for challenging times*. Cambridge, Massachusetts: Harvard Education Press.

Torgerson, W. S. (1960). *Theory and Methods of Scaling. New York:* John Wiley.

Tyack, D. (1974). *The One Best System: A History of American Urban Education* (Cambridge: Harvard University Press.

U.S. Department of Education (November 2009). Race to the Top: Executive Summary. Washington, DC: US Department of Education. Retrieved from http://ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education (March 2010). *A blueprint for reform. The reauthorization of the Elementary and Secondary School Act*. Washington, DC: US Department of Education. Retrieved from http://www2.ed.gov/policy/elsec/leg/blueprint/.

Zastrow, C. & Janc, H. (2004). *Academic atrophy: The condition of the liberal arts in America's public schools.* Washington DC: Council for Basic Education.