International Journal of Education Policy & Leadership

The Unintended, Pernicious Consequences of "Staying the Course" on the United States' No Child Left Behind Policy

Audrey Amrein-Beardsley
Arizona State University

The phrase "no child left behind" has become a familiar expression in American education circles and in popular culture. The sentiment implied by these four words is noble. However, the effects of the top-down implementation of the high-stakes testing provisions of the law have been anything but salutary for public school children, teachers, and administrators. This claim is supported by data describing many of the ways in which well-intentioned but desperate educators, from the statehouse to the schoolhouse, have been driven to game the system in ironic defense of the children, teachers, and administrators least equipped to defend themselves. It is argued herein that, instead of reauthorizing the stronger accountability tenet of NCLB, it might do very well to let it fade away.

Amrein-Beardsley, A. (2009). The Unintended, Pernicious Consequences of "Staying the Course" on the United States' No Child Left Behind Policy. *International Journal of Education Policy and Leadership 4*(6). Retrieved [DATE] from http://www.ijepl.org.

Introduction

> I want to thank Secretary Spellings and her fine team for welcoming me here to the Department of Education. I have just reassured the Secretary and the folks who work here that the reauthorization of the No Child Left Behind Act is a priority of this administration. And the reason I say it's a priority is because this act is working. We strongly believe in setting high standards for all students, and we strongly believe that, in order to make sure those standards are met, we must measure to determine whether or not the schools are functioning the way we expect them to function, and the way the parents expect them to function, and the way the taxpayers expect them to function.
>
> —President George W. Bush, NCLB Reauthorization Speech, October 6, 2006

President Bush and other educational leaders believe that setting high standards and holding students accountable for meeting high standards are the foundation of educational reform. Attaching incentives to learning and sanctions to poor performance is assumed to increase student achievement by motivating students to learn more, teachers to teach more effectively, and administrators to implement better educational programs. According to this line of logic, as written into the United States' No Child Left Behind (NCLB) policy in existence since 2002, the best way to promote student learning and achievement is to reward and penalize students, teachers, administrators, schools, and school districts according to student performance on standardized tests.

Tests used to grant or deny high school diplomas, also known as graduation or exit exams, are the most common high-stakes accountability devices across states For individual students, high test scores might bring about college scholarships, exceptional academic awards, or marks of distinguished achievement, while low test scores may bring about retention in grade level or the denial of a high school diploma. For teachers, high test scores might bring about financial bonuses or increases in salary, while low test scores may cause a teacher to be fired or transferred to a different school. For administrators, high test scores might warrant cash bonuses and low test scores might result in administrative transfers, contract termination, or job loss. For schools and school districts, high test scores might merit monetary awards and public approval, while low test scores may bring about public criticism and school

ASCD Simon Fraser University George Mason University
This peer reviewed journal is a joint publication of the Association for Supervision and Curriculum Development (ASCD), the Faculty of Education at Simon Fraser University (SFU), and the College of Education and Human Development at George

reconstitution or closure. Children attending low-performing schools may also apply for transfer to another public school with higher test scores, as written into NCLB.

But are the high standards and accountability components written into NCLB in fact working as President Bush suggests? Does the record show that raising academic standards and attaching serious consequences to tests encourage students to learn and achieve more?

Some researchers have provided evidence that high standards and accountability have increased student achievement (Braun, 2004; Carnoy & Loeb, 2002; Rosenshine, 2003; Scheurich, Skrla, & Johnson, 2000; Schiller & Muller, 2000), and others have provided evidence refuting this claim. This second group of researchers argues that states that have implemented strong, standards-based accountability policies—regardless of their level of punitiveness—have advanced no farther than states that have not implemented such measures (Amrein-Beardsley & Berliner, 2002b,c; Amrein-Beardsley & Berliner, 2003; Camilli, 2000; Klein, Hamilton, McCaffrey, & Stecher, 2000; Nichols & Berliner, 2008; Nichols, Glass, & Berliner, 2005; Haney, 2000; Heubert & Hauser, 1999; Marchant & Paulson, 2005; Marchant, Paulson, & Shunk, 2006). In fact, these researchers (see also Amrein-Beardsley & Berliner, 2002a; Haney, 2001) and others (Clarke, Haney, & Madaus, 2000; Kohn, 2000; McNeil, 2000; Sacks, 1999) suggest that, if anything, highly punitive accountability strategies may be producing unintended negative consequences, the gravity of which outweighs whatever benefits these policies might promote.

When former Texas governor Bush first ran for president in 2000, he brought with him what he declared was *the proof* that standards linked to accountability mechanisms increase student learning and achievement. In Texas, large gains in National Assessment of Educational Progress (NAEP) scores were heralded as evidence that these policies did, undeniably, improve student learning and achievement. Before these state level policies were implemented, levels of student achievement in Texas were far below the national average; after implementation, Texas children's test scores surpassed the rest of the nation (Grissmer, Flanagan, Kawata, & Williamson, 2000).

Graphs illustrating these phenomenal gains were celebrated, and the sensational claim was termed "the Texas Miracle" (for more information, see Bracey, 2008; Haney 2000). Governor Bush's high standards and accountability policies apparently produced miraculous effects in Texas and, he reasoned, would undoubtedly increase performance of all schools throughout the nation. This is how he justified his proposed education policies during the presidential campaign of 2000, and he used these claims to justify his aspiration to become known as the Education President.

During this time, however, researchers who reexamined these phenomenal gains in student achievement (Amrein-Beardsley & Berliner, 2002c; Haney, 2000) found that the gains in Texas weren't all they were cracked up to be. At the same time that children in Texas posted amazing gains, the percent of students excluded from participating in the NAEP increased at extraordinary rates. Paradoxically, the nation's exclusion rate (outside Texas) declined at the same time.

Haney (2000) described this as an "illusion arising from exclusion" because significantly more students with disabilities and non-English speakers were excluded from participating in the Texas NAEP. Whether the phenomenal gains posted in Texas were a product of authentic learning or should be attributed to fewer low-performing students participating became the subject of a fierce dispute, both scientific and political. This ultimately divided people into two camps: Those who still believed a miracle had occurred in the Texas education system and those who believed the miracle was more mythical, contrived through inflated and selective student exemption practices. Critical researchers continue to question whether more testing increases student achievement and whether reported test score increases reflect improved student learning or multiple methods of "gaming the system."

Gaming the System

In *The Prince*, Machiavelli makes cunning recommendations to the prince so that he might game the system and secure prince-hood, public support, control of the people, and aristocratic power. Machiavelli's tactical suggestions to the prince include: justify human extermination, eliminate public freedom, embrace opportunism, destroy resistance, seduce the public, and mask evil with good. He argued that extreme, cruel, deceitful, immoral, and unethical measures are warranted and necessary to pursue power, the ultimate end for the prince.

Machiavelli was eventually tortured, imprisoned, and exiled because of his brutally honest book that only thinly disguised the disclosures of the inner workings of politics in renaissance Italy. Yet Machiavelli continues to help us understand how people game social systems for personal benefit, power, and control.

People who game the system understand the rules, policies, and procedures of the game and are equipped to manipulate and take advantage of loopholes in the system. In fact, gaming the system is easiest when rules, policies, and procedures are imprecise and ambiguous, permitting the game to continue. Unfortunately through exploiting loopholes in these rules, educators may compromise higher values: honesty, integrity, and worthiness of public trust, to name a few. In true Machiavellian form, people who game the system compromise their morals and ethics to achieve self-justified ends.

Instances of gaming the system are widespread and permeate all social-political systems. Some lawyers game the system for their guilty clients by exploiting the insanity plea. Some highly sought accountants are proficient at gaming the system by finding or creating questionable tax shelters, breaks, exemptions, deductions, and the like. Enron is the paradigmatic case of gaming the system in big business. What about those who game the system by illegally downloading, justified as *legally borrowing*, movie and music electronic files from strangers who allegedly want to share their files via Limewire.com? Others game the system by using radar detectors to evade speeding tickets. Retailers game the system by overpricing store items and then advertising them at 50 percent off the (artificially inflated) price. Physicians game the system by overcharging insurance companies for their services and justify their fraud because insurance companies are simultaneously gaming the insured.

All social systems have players in the game, one way or another, and the only way to control these players is to close the loopholes perpetuating such gaming or eliminate the practices causing such gaming altogether. In policy analysis, Machiavelli can help us better understand the ways players game such policies (Radin, 2000).

Gaming the System to Meet NCLB's 100% Proficiency Target

NCLB requires all states to implement accountability policies to ensure that 100 percent of elementary students in grades 3–8 and high school students in public schools achieve academic proficiency by the year 2014. Every public school student in the nation is to reach academic proficiency eight years from now. Will 100 percent of America's public school students reach this target? By gaming the system, it is likely.

Teachers, administrators, and education leaders have employed a multitude of questionable test preparation practices to help their states, schools, and students meet high standards. Methods of gaming tests, however, result in spurious test score gains unrelated to true gains in student learning. When investigating whether stronger accountability measures help students meet higher standards, we must consider the extent to which the following factors are used to artificially inflate gains in student learning and academic achievement.

Teaching to the Test

Teaching to the test occurs when teachers disproportionately teach students things they know will be on accountability tests. A teacher who has administered a few of these annual tests in the past may gain some understanding of what to expect and teach students only those concepts the teacher predicts will be on future iterations of the test. A teacher may rehearse students for a test with clone items that look exactly like the items on previous forms of the test but with the names of the people in the word problems and the numbers in the mathematical equations changed. A teacher may have students write and rewrite five-paragraph essays, neglecting other writing genres, knowing that a five-paragraph essay is expected on the annual writing assessment. A teacher may make copies of the actual test or the test used in previous years to rehearse students for the upcoming tests, over and over again. Teachers might have their students spend hours memorizing facts, learning test-taking strategies, bubbling score sheets accurately, eliminating unlikely distractor responses, making educated guesses, and using multiple-choice answers to solve mathematical problems backwards, all of which help students game these tests to pass; all of which are classic threats to test validity.

Because teaching to the test may cause scores to increase, it is a popular practice in which teachers engage to artificially raise test scores. Such practices are sometimes even encouraged by local school administrators when school composite statistics are at risk. Score gains do not last, however, nor are they reflected in other measures of student learning and achievement (See, for example, Amrein-Beardsley & Berliner, 2002b,c; Heubert & Hauser, 1999; Linn, Graue, & Sanders, 1990; McNeil, 2000; Stake, 2001).

Narrowing of the Curriculum

Narrowing the curriculum is when teachers do not teach some important topics within subject areas or avoid teaching parts of the state standards they are supposed to teach, knowing that what they omit from their lessons

will not be included on accountability tests. It may be written in the state standards, for example, that a 10th grade mathematics teacher must teach graphing equalities and inequalities. A teacher aware of the fact that questions assessing students' abilities to graph inequalities are usually not included on the high school graduation exam might simply omit this lesson to concentrate more on graphing equalities instead.

School administrators also contribute to this at the local school level. Two months before high-stakes tests are administered, a school principal may eliminate recess, art, music, or physical education or replace science with mathematics and social studies with language arts to intensify math, reading, and writing instruction to provide amplified opportunities for students to rehearse the basic subject areas tested (See, for example, Dorn, 1998; Koretz, 1996; Kreitzer, Madaus, & Haney, 1989; McNeil, 2000; Sacks, 1999; Swope & Miner, 2000.)

Exclusion and Exemption Practices

Students are also subjected to creative exclusion and exemption practices. Students with histories of poor academic performance might be encouraged to stay home and miss accountability tests, or they might be suspended or expelled before accountability tests are administered. Low-scoring high school students might be counseled to quit or be suspended from school just before tests so that their scores will not be included in composite test score calculations. Students may be falsely exempted from participating in accountability tests for being English language learners (ELLs) even if they speak English fluently enough to participate. Additionally, students may be purposely labeled as severely handicapped when, by law, their handicap should not prevent them from participating in state tests. (Federal and state provisions have been enacted, however, to minimize these false exemptions of ELLs and special needs students.) Low-performing students may also be retained in grade levels in excessive numbers before pivotal testing years so that they will have more chances to be drilled on the tested material or so that they will not taint the pool of test takers by negatively skewing test score distributions. School personnel would rather these students not take part in accountability tests. In all probability, if these students participated, they would bring down the school's average scores, placing the district, school, administrators, and teachers at risk (See, for example, Bass, Dizon, & Feller, 2006; Haladyna, Nolen, & Haas; 1991; Haney, 2000; Heubert & Hauser, 1999; Kelleher, 1999; Klein, Hamilton, McCaffrey, & Stecher, 2000; Madaus & Clarke, 2001; Madaus, West, Harmon, Lomax, & Viator, 1992; McGill-Franzen & Allington, 1993; McNeil, 2000; May, 2000).

The Bubble Kids

School personnel have also refocused energies on bubble students, also known as *borderline students*, who are on the border of passing or failing high-stakes tests. Because these students are more likely than their lower scoring peers to post passing scores, school personnel often focus inordinately and intensively on these students to help them acquire the knowledge necessary to pass these tests. Each passing score posted by a borderline student translates into an increase in overall test averages and, more importantly, student proficiency percentages. Students above the borderline guarantee good scores, so they are left alone; students below the borderline, for whom school personnel have the least amount of hope, go about the normal school day, neglected because they are least likely to pass the tests or contribute to increased averages and proficiency percentages (Kohn, 1999; Madaus, West, Harmon, Lomax, & Viator, 1992; McNeil, 2000; Schrag, 2000).

Cheating

The pressures associated with stronger accountability testing are also driving teachers and school administrators to cheat. In fact, as tests become more consequential and the penalties of failure more severe, the likelihood school personnel will cheat on tests increases (Schrag, 2000; Viadero, 2000). In newspaper articles across the country, journalists have described ways in which teachers and administrators have cheated on accountability tests; the articles written because they have been caught. A teacher may allow students more time to complete a test than is prescribed; walk around the classroom providing students with hints, clarifications, definitions, or answers; tell students to rethink particular questions if the teacher sees incorrect answers; and some have been caught manually correcting students' answers on accountability test score sheets. Cheating is just one more way teachers and administrators can artificially promote increases in test scores to dodge the negative or realize the positive consequences attached to accountability tests (Haladyna, Nolen, & Haas, 1991; Haney, 2000; Kornhaber & Orfield, 2001; Sacks, 1999; Shepard, 1990; Smith, 1991; Urdan & Paris, 1994).

Administrative Manipulation

Administrators are not immune to the temptations of cheating and other score-boosting practices. Administrators have briefed teachers on what will be tested on upcoming accountability tests; made copies of secure tests and distributed them to teachers before official tests are administered; and changed low-scoring students' identification numbers to make their score sheets invalid, resulting in the exclusion of their scores from the school's composite statistics. Administrators may hire test-boosting consultants who encourage teachers to focus instruction only on those students who they feel have a fighting chance of passing accountability tests or initiate mass exoduses of low-scoring students who will do nothing for composite test reports but bring school results down. Conversely, administrators might encourage students who have already posted high scores on accountability tests to participate every year to boost overall school averages. Administrators may use funds—even entire textbook budgets—to purchase test preparation booklets filled with test practice worksheets guaranteed to boost test scores, provided that children are rigorously drilled on one test prep activity after another. Administrators may also narrow the curriculum by concentrating all personnel efforts on the subject areas "that matter," after which significant gains in scores are celebrated, only to realize significant drops in achievement in the neglected subject area(s) at the same time (See, for example, Booher-Jennings, 2006; Gordon & Reese, 1997; Goodnough, 2001; Kohn, 2000; Madaus, West, Harmon, Lomax, & Viator, 1992; McNeil, 2000; McNeil & Valenzuela, 2001; Nichols, Glass, & Berliner, 2005; Schrag, 2000; Smith, 1991).

Dumbing Down Tests and Manipulating Cut Scores

Almost always when state accountability tests are first administered, extremely high rates of student failure are published in state newspapers. The public is sent into a frenzy, and politicians and the public usually blame teachers, administrators, and low standards and expectations for the lack of student proficiency. In actuality, however, the initial high rate of student failure can be better explained by two factors: (1) How difficult and unfamiliar the accountability test was, and (2) Where the pass/fail cut score was set. Because it is not politically feasible to fail too many students year after year, over time accountability tests are made easier, cut scores are lowered, and more students pass. This gives the public the false impression that, because of initial high failure rates, the threat of accountability tests worked: the threat of sanctions motivated students to learn more, teachers to teach more effectively, and administrators to adopt better educational programs. In actuality, what happens behind the scenes in state departments of education and legislative committees often has a more significant role in generating apparent gains in student achievement. Dumbing down tests and manipulating cut scores are two more ways to manufacture increased levels of student proficiency and politically acceptable pass rates (See, for example, Haney, 2000; Kellow & Wilson, 2001; Koretz, Linn, Dunbar, & Shepard, 1991; Madaus & Clarke, 2001; Orfield & Kornhaber, 2001; Rudner, 2001; Schrag, 2000).

A Case in Point

In 1996, Arizona State Board of Education members decided to develop new, criterion-referenced competency tests in math, reading, and writing to hold children accountable for meeting state standards in grades 3, 5, 8, and 10. All students attending an Arizona public school were to be assessed on the new Arizona Instrument to Measure Standards (AIMS) tests. Students in grade 10 were required to take and pass all three subtests of the grade 10 AIMS to receive a high school diploma. If students did not pass the graduation test on their first attempt, they could continue to take the subtest(s) they did not pass until they met or exceeded the state standards in all three areas. They were permitted to retake failed tests twice annually until age 22.

After the tests were first administered in 1999, however, complaints arose that the tests were too difficult. After students failed the AIMS at astonishing rates, particularly in math, state officials made the AIMS tests easier. In 2000, members of the Arizona Board decided to remove all short-answer, constructed response items previously included on the math and reading AIMS tests—the revised test would become 100 percent multiple choice. They also removed more difficult items and reduced the length of the math test.

Complaints also arose that the some of the content tested was invalid. For example, the 10th grade math test included trigonometry problems, but trigonometry was not taught until the 11th grade. How could Arizona's students be held accountable for knowing trigonometry that they had not yet had the opportunity to learn? So the state made the tests more valid through improving alignment with standard curricula, which consequently made the tests easier again. At the same time, the state lowered

the passing grade, or cut score, allowing more students to pass the high school graduation exam.

Initially, the class of 2001 was to be the first class to either pass the AIMS high school graduation exam or be denied high school diplomas. But implementation of this provision was postponed for the class of 2002; in 2001, implementation was officially pushed back again to first take effect for the class of 2006.

However, after tens of thousands of students continued to struggle to pass the AIMS, educational leaders at all levels began gaming the system and creating additional loopholes to help students pass the tests, particularly in anticipation of the high rates of student failure which were predicted and the 100 percent proficiency target written into NCLB (Kossan, 2006; Yara, 2006).

State officials gave teachers better guidelines on what to expect on the tests; rewrote state standards to indicate which components of the state standards would or would not be tested; and uploaded practice tests with clone items onto the Arizona State Department of Education Web site for teacher, parent, and student access and practice. All of these mechanisms were provided with the good intentions of helping teachers teach the standards and administrators align curricula more effectively. But they were often, and continue to be, misused as they enable many to teach to the test and narrow the curriculum more efficiently and accurately.

In 2004 nine Arizona school districts were accused of "engaging in illegal actions to help students appear to have learned more than they had in actuality" on the AIMS tests (Haver, 2004). In one case, a principal of a school in one of the largest school districts in the Phoenix metropolitan area had to defend herself before the school board, pleading that she did not improperly alter students' test scores. She resigned and continues to deny the allegations (Haver, 2004). Known instances of cheating in Arizona are relatively minor, though, when compared to reports of cheating in states with more dramatic and immediate consequences riding on students' performance on high-stakes accountability tests (Nichols, Glass, & Berliner, 2005).

In 2005 the AIMS tests were altered again to assess the new state standards adopted by the Arizona State Board of Education in 2003 and to include criterion-referenced and norm-referenced test items. Because Arizona state statute requires the state to administer criterion-referenced and norm-referenced tests as part of its state assessment program, state officials wisely commissioned a new assessment to serve both purposes at once. By including these test items in the revised AIMS tests, they cut the time it took to administer the two state-mandated tests in half while continuing to provide the state with the test data reports required by law. It was a win-win situation. Nonetheless, it meant implementing yet another version of the AIMS and establishing another set of cut scores.

The AIMS Dual Purpose Assessment (DPA) was presented and approved by members of the Arizona Board in 2005. The AIMS DPA, the test currently used in Arizona, includes a combination of criterion-referenced items taken from the AIMS and norm-referenced items taken from the TerraNova, a nationally normed test created by CTB/McGraw-Hill. The new tests were also expanded to assess levels of student learning continuously in grades 3–8 and in high school. Likewise, new benchmarks for meeting the state standards were set. As expected, the second significant revision to the AIMS tests caused marked increases in levels of student achievement.

When students in the class of 2006 first took the AIMS test as sophomores in 2004, 39 percent passed the mathematics AIMS, 59 percent passed the reading AIMS, and 62 percent passed the writing AIMS. The public was dismayed, and state officials and the public blamed teachers and administrators for not doing their jobs and blamed children and their families for not meeting the higher standards adopted by the state. Some of these criticisms may have had merit, but the poor levels of student achievement could have probably been better explained by the difficulty of the tests and the positioning of the cut scores at the time.

In contrast, when students in the class of 2007 first took the revised AIMS DPA as sophomores in 2005, 69 percent of the sophomores passed the math section (a 30 percent increase in one year); 75 percent passed reading (a 16 percent increase in one year); and 74 percent passed writing (a 12 percent increase in one year) (Kossan, Ryman, & Konig, 2006). Miraculous gains in achievement were reported, which were mostly—if not entirely—due to the administration of the new and significantly easier AIMS DPA. Nonetheless, educators, members of the media, and politicians commended teachers and administrators for doing their jobs so much better and congratulated school children for taking their learning and achievement more seriously. It is not that these good things did not occur, but little mention was made of the major alterations to the standards, the tests, the cut scores, and the lowering of the bar.

One year later, to guarantee that still more Arizona students would meet the higher standards set forth by

the state, the Arizona school superintendent wrote that he was going to make it his mission to help more students pass the AIMS tests. He promised to do this by ensuring that teachers, parents, and students knew what was on the test, by developing and administering more test-preparation workbooks, and by providing free tutoring to students in need (Fisher, 2006). The latter option represented the only educationally defensible test preparation practice (Popham, 2004) to increase student learning. That is, if students were tutored on the standards and *not only* on test-taking tactics, content-focused tutoring would be a legitimate way to increase student learning. State legislators allocated $10 million to tutor juniors and seniors who hadn't passed the test, and 6,000 of 37,000—or 16 percent—eligible students took advantage of the program (Roberts, 2006).

One high school designed classes around AIMS to help seniors past the test or particular portions they had failed (Yara, 2006). These classes, titled AIMS Math, AIMS Reading, and AIMS Content, were offered in Triage Academies in which students rehearsed for the AIMS two hours per day (Kossan, Ryman, & Konig, 2006). At another school, an after-hours AIMS homework club was developed to help high scorers earn college scholarships and low scorers improve their chances of passing the AIMS (Homework Clubs, 2006). At another school, students were touted as students who "loved to write" as evidenced by their mastery of the six traits of writing, a rubric that happens to be the scoring guide used to assess all student writing samples on the AIMS writing test (Madrid, 2006).

As Figure 1 illustrates, it is difficult to differentiate between increased levels of student learning and system-gaming influences to explain the dramatic jumps in test scores because they immediately follow state policy measures to make the tests easier and lower passing scores. It is also impossible to determine how other methods of gaming the system applied in local school settings further artificially increased levels of student achievement throughout this period. Publicly available data, downloaded from the Arizona Department of Education Web site (http://www.ade.az.gov), represents the percent of students who met or exceeded Arizona state standards on the grade 3, 5, 8 and 10 math AIMS, and they are presented here.

Figure 1 (page 13) illustrates points at which accountability tests were made easier or cut scores were lowered. But what is impossible to capture in these trend lines are the cases in which teachers and administrators may have gamed the system in more invisible, subtle, yet significant ways previously discussed and too frequently used to artificially boost students' test scores. It is not possible to capture phenomena like teaching to the test, narrowing of the curriculum, instances of cheating (especially since cheating is under-reported), test coaching, indefensible test preparation practices, test drills and rehearsals, using clone items to prepare students for tests, and student exemptions and expulsions. However, we know such methods of gaming the system persist, and we know they work to artificially boost student test scores. These methods bloom when children, teachers, administrators, schools, and districts are to be penalized for poor or rewarded for solid test performance and when consequences attached to student performance on accountability tests become more consequential.

For clarity, suggest changing to: In 2006 after last-ditch efforts to postpone using the test as a graduation requirement failed, the state of Arizona, in a final effort to prevent a large number of students from failing the AIMS test, allowed thousands of high school students to apply bonus points from the grades they earned of *C* or higher in core classes to trump their failing AIMS scores. In 2006, of the 64,000 students who started the school year as seniors (the first class to be held accountable for their high-stakes test performance), 14,000-plus dropped out or moved. Of the remaining 50,000 seniors, 46,111 (92.2 percent) passed AIMS; 2,630-plus (?5.3 percent) failed AIMS but passed using bonus points from core classes; and 297 (0.6 percent) were denied a high school diploma (Kossan, 2006). In the end, this was quite a trivial number, especially given the complex, aforementioned history and the overarching goals and intentions of the policy.

When asking whether high-stakes tests help students meet higher standards, we must take care to discriminate between factors promoting authentic gains in learning and achievement and the multitude of ways to manufacture artificial student test score gains. Otherwise, manufactured student achievement scores and tweaked accountability tests may ultimately help states meet the 100 percent proficiency target written into NCLB but will likely do little to support or advance genuine student learning.

These methods of gaming the system need to be considered when differentiating between true and falsified levels of increased student learning and achievement. And when interpreting trend line, longitudinal data consumers need to be aware that many extraneous, uncontrollable variables might help explain gains or "illusions of progress," when true gains in student learning cannot

be illustrated alone. This was exemplified in this case. There was much occurring behind the scenes that helped explain student achievement in Arizona.

And even though gains in levels of student achievement cannot be explained solely as effects of gaming the system, such gains should not be attributed solely to policies and practices meant to hold students accountable for meeting higher standards. At the very least, we might agree that some proportion of the claimed successes of the testing provisions of NCLB are artificial as manufactured by teachers, administrators, and local and state education leaders who today, as foreshadowed in the case of the Texas Miracle, are continuing to work the system to survive.

Glimpse of a Silver Lining

In spite of the many unintended, negative consequences resulting from confusing testing, with more effective teaching, the standards-based reform movement driven by NCLB may have promoted a few forms of educational improvement. In cost-benefit terms, however, the few benefits have occurred at a cost—and a significant monetary cost considering the billions of dollars spent purchasing standardized tests and their accompanying test preparation materials (profits for which have increased exponentially since the enactment of NCLB). Ironically, test developers are also gaming the system, but to their advantage.

Perhaps the largest benefit associated with high-stakes testing policies is that states have revisited, revised, and raised their standards to meet what professional teacher organizations (e.g., the National Council of Teachers of Mathematics) regard as essential subject matter knowledge. Because teachers are being held accountable for teaching the content included in these standards, the result has been more consistency across subjects and better uniformity across classrooms, schools, and even state borders. However, not everyone in the education profession agrees this is a benefit. Along with an increase in standardization and uniformity has come a decrease in professional autonomy for teachers, administrators, and local state school personnel.

Another benefit of implementing high-stakes testing policies is that more monies are being targeted toward some children most in need of help. Remediation programs have been developed to help students who fail high-stakes tests learn the material necessary to pass each test. Although teaching children how to pass high-stakes tests is of questionable utility and principle, these efforts are being targeted toward students who, it is assumed, went without such remediation efforts before NCLB.

A final benefit of nearly universal high-stakes testing may be that test scores can be used for diagnostic purposes, helping teachers better understand child learning problems and design individualized instruction. However, in most states, high-stakes tests are administered in the spring and the test results are returned during the summer, after school is out. Children have exited last year's classrooms to start a new school year in the fall with a new teacher. To date, the diagnostic benefits of implementing high-stakes tests have gone unrealized and, ironically, may only be attained as greater numbers of children are held back in grade level, which is a politically unacceptable option.

Concluding Remarks

When rational people are faced with impossible demands, they will work every angle they can to adapt, survive, and possibly prosper. Rational people are engaging in irrational, unethical, and unprofessional practices as they attempt to dodge the negative or realize the positive consequences increasingly tied to student performance on tests. Many educators faced with the stronger accountability provisions written into NCLB are working every angle to meet the letter of the law, and ironically, through their practices, are working against the well-intentioned theory on which the law is based. Although there are educators who, despite the temptation to game the system, are following the law in ethical and professionally defensible ways, this paper described the lengths to which many educators at local and state levels have gone to ensure that students do well on high-stakes tests when focusing on student learning isn't enough. The hands of these good-hearted public educators are tied by a misguided federal law, and bad things are happening.

If federal lawmakers' continue to believe that the more educators and students are held accountable and the more stringent the consequences, the more student test scores will increase but student learning won't. As long as methods of gaming the system persist, never will we be able to determine precisely what students genuinely learn under NCLB. But with serious consequences attached to performance, it is unlikely methods of gaming the system might ever be eliminated or controlled. Other tactics might simply emerge, again, to counter the high-stakes, high-stress environment now encapsulating student learning in America's public schools.

The only way to control these players is to close the loopholes perpetuating such gaming. With the reauthorization of NCLB forthcoming, ways educators are gaming NCLB might be understood and used to inform significant changes in the future version of this law.

Or, it might be preferable to eliminate the practices causing such gaming altogether: The stronger accountability tenet at issue here might be discarded. Nichols and Berliner (2008) detail why they believe high-stakes tests have so easily slipped into contemporary American life. They argue that (1) it makes sense to the general public to apply basic business principles to education in terms of monitoring inputs and, more importantly, outputs; (2) in order to maintain global competitiveness, students must be held accountable for meeting even higher standards; (3) this type of sorting mechanism helps preserve social status by effectively differentiating the most from the least capable, as measured by standardized tests; (4) the parent(s) of the students least effected by high-stakes testing are least concerned yet the most able to take action against it; and (5) the competition-seeking public intrinsically "likes" high-stakes testing. By educating the general public about how high-stakes tests have contaminated student learning in America's public schools and by helping them put the above assumptions "in check" and collectively question whether what seems like commonsense indeed works in the ways theorized, the stronger accountability tenet of NCLB might very well be abandoned versus reauthorized altogether.

In Machiavelli's later work, *Discourses on Livy*, he promotes republicanism—the governing of a nation through liberty, civic virtuosity, and popular consent and control. Machiavelli valued liberty most during times of oppression and, in his later years, was more interested in promoting just and noble ends for the popular good. Perhaps the nation might follow a similar trajectory with the reauthorization of the stronger accountability tenet of NCLB nearing. Perhaps the people, tired or fed up with the uniform, oppressive nature of this tenet of the law, may very well become more interested in promoting just and noble ends for the popular good; that is, promoting quality education and bringing back to public schooling the liberty and civic virtuosity it deserves and on which it thrives.

References

Arizona Department of Education, Assessment Section. (2005). *AIMS Update*. Retrieved May 16, 2005, from http://www.azed.gov/standards/aimsupdates/2004-2005/6-2005.pdf

Arizona Department of Education Standards and Assessment Division. (2002, August 21). *Press release summary report for spring 2002 test administration: For grades 3, 5, 8, and high school reading, writing, and mathematics*. Retrieved October 20, 2004, from http://www.ade.state.az.us/ResearchPolicy/AIMSResults/2002AIMSExecSum.pdf

Amrein, A. L. & Berliner, D. C. (2002a). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18). Retrieved September 18, 2008 from http://epaa.asu.edu/epaa/v10n18/

Amrein, A. L. & Berliner, D. C. (2002b). The impact of high-stakes tests on student academic performance. Research funded by the Great Lakes Center and published by the Educational Policy Research Unit.

Amrein, A. L. & Berliner, D. C. (2002c). *An analysis of some unintended and negative consequences of high-stakes testing*. Research funded by the Great Lakes Center and published by the Educational Policy Research Unit.

Amrein-Beardsley, A. & Berliner, D.C. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives*, 11(25). Retrieved September 18, 2008 from http://epaa.asu.edu/epaa/v11n25/

Bass, F., Dizon, N. Z., & Feller, B. (2006, April 18). *States help schools dodge No Child: Accused of 'gaming the system.'* Chicago Sun-Times.

Booher-Jennings, J. (2006, June). Rationing education in an era of accountability. *Phi Delta Kappan*, 87. Retrieved November 12, 2006, from http://www.pdkintl.org/kappan/k_v87/k0606boo.htm

Bracey, G. W. (2008). Kicked down and out by the Texas Miracle. *Phi Delta Kappan*, 89(9), 699–700.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Educational Policy Analysis Archives*, 12(1). Retrieved November 8, 2006, from http://epaa.asu.edu/epaa/v12n1/

Camilli, G. (2000). Texas gains on NAEP: Points of light? *Education Policy Analysis Archives*, 8(42). Retrieved November 12, 2006, from http://epaa.asu.edu/epaa/v8n42.html

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305–331.

Clarke, M., Haney, W., & Madaus, G. (2000, January). High-stakes testing and high school completion. *NBETPP Statements*, 1(3). Retrieved November 8, 2006, from http://www.bc.edu/research/nbetpp/statements/V1N3.pdf

Dorn, S. (1998, January 2). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1). Retrieved November 12, 2006, from http://olam.ed.asu.edu/epaa/v6n1.html

Firestone, W. A., Monfils, L., Camilli, G., Schorr, R. Y., Hicks, J. E., & Mayrowetz, D. (2002). The ambiguity of test preparation: A multimethod analysis in one state. *Teachers College Record*, 104(7), 1485–1523.

Fisher, G. (2006, May 6). AIMS test needs to go, not come back. *The Arizona Republic*.

Goodnough, A. (2001, June 14). Strains of fourth-grade tests drives off veteran teachers [Electronic version]. *New York Times*. Retrieved November 12, 2006, from http://www.nytimes.com/2001/06/14/nyregion/14GRAD.html

Gordon, S. P., & Reese, M. (1997). High-stakes testing: Worth the price? *Journal of School Leadership*, 7, 345–368.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP test scores tell us* [Electronic version]. Santa Monica, CA: RAND Corporation. Available: http://www.rand.org/publications/MR/MR924

Haladyna, T., Nolen, S. B., & Haas, N. S. (1991). Raising standardized test scores and the origins of test score pollution. *Educational Researcher*, 20(5), 2–7.

Haney, W. (2000). The myth of the Texas Miracle in education. *Education Analysis Policy Archives*, 8(41). Retrieved November 8, 2006, from http://epaa.asu.edu/epaa/v8n41/

Haney, W. (2001). *Revisiting the myth of the Texas miracle in education: Lessons about dropout research and dropout prevention*. Paper prepared for the Dropout Research: Accurate Counts and Positive Interventions Conference, Cambridge, MA. Retrieved November 8, 2006, from http://www.skirsch.com/politics/education/RevisitingTXMyth.pdf

Haver, J. (2004, September 13). Cheating on AIMS tests: It's not students doing it. *The Arizona Republic*.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation* [Electronic version]. Washington, DC: National Academy Press. Retrieved November 8, 2006, from http://www.nap.edu/html/highstakes/

Homework clubs cool way to study, pass the AIMS test. (2006, March 24). *The Arizona Republic*.

Kelleher, M. (1999, June). Dropout rate climbs as schools dump truants. *Catalyst Chicago*. Retrieved September 18, 2008 from http://www.catalyst-chicago.org/news/index.php?item=330&cat=23

Kellow, J. T., & Wilson, V. L. (2001). Consequences of (mis)use of the Texas Assessment of Academic Skills (TAAS) for high-stakes decisions: A comment on Haney and the Texas miracle in education. *Practical Assessment, Research & Evaluation*, 7(24). Retrieved September 9, 2008, from http://PAREonline.net/getvn.asp?v=7&n=24

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B.M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, 8(49). Retrieved November 8, 2006, from http://epaa.asu.edu/epaa/v8n49/

Kohn, A. (1999). *The schools our children deserve: Moving beyond traditional classrooms and "tougher standards."* New York: Houghton Mifflin Company.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

Koretz, D. M. (1996). Using student assessments for educational accountability. In E. A. Hanushek and D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives*. Washington, DC: National Academy Press.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The effects of high-stakes testing an achievement: Preliminary findings about generalization across tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

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

Kossan, P. (2006, July 27). In-class work lets many pass AIMS test. *The Arizona Republic*.

Kossan, P., Ryman, A., & Konig, R. (2006, July 13). More kids pass '05 AIMS: Questions raised of easiness of test vs. real gains. *The Arizona Republic*.

Kreitzer, A. E., Madaus, G. F., & Haney, W. (1989). Competency testing and dropouts. In L. Weis, E. Farrar, & H. G. Petrie (Eds.), *Dropouts from school: Issues, dilemmas, and solutions*. Albany, NY: State University of New York Press.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that 'everyone is above average.' *Educational Measurement: Issues and Practice*, 9(3), 5–14.

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. A. (1992). *The influence of testing on teaching math and science in grades 4–12*. Chestnut Hill, MA: Center of Study of Testing, Evaluation, and Educational Policy, Boston College.

Madrid, O. (2006, February 5). Pupils absorb 6 traits on way to mastering art of writing: Basic components of composition taught in preparation for AIMS test. *The Arizona Republic*.

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 November 8, 2006, from http://epaa.asu.edu/epaa/v13n6

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 November 20, 2006, from http://epaa.asu.edu/epaa/v14n30/

May, M. (2000, October 4). State fears cheating by teachers: 51 schools left off cash award list. *The San Francisco Chronicle*. Retrieved November 12, 2006, from http://www.sfgate.com

McColskey, N., & McMunn, N. (2000). Strategies for dealing with high-stakes state tests. *Phi Delta Kappan*, 82(2), 115–120.

McGill-Franzen, A., & Allington, R. L. (1993). Flunk'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher*, 22(1), 19–22.

McNeil, L. (2000). *Contradictions of school reform*. New York: Routledge.

McNeil, L., & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Nichols, S. L., & Berliner, D. C. (2008). Why has high-stakes testing so easily slipped into contemporary American life? *Phi Delta Kappan*, 89(9), 672–676.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2005, September). *High-stakes testing and student achievement: Problems for the No Child Left Behind Act*. Tempe, AZ: Arizona State University, Education Policy Studies Laboratory. Retrieved November 8, 2006, from http://epicpolicy.org/files/EPSL-0509-105-EPRU.pdf

Office of the Press Secretary. (2006, October 5). *President discusses NCLB reauthorization at the education department*. Retrieved November 8, 2006, from http://www.whitehouse.gov/news/releases/2006/10/20061005-4.html

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

Popham, W. J. (2004). *Classroom assessment: What teachers need to know*. Boston: Allyn & Bacon.

Radin, B. (2000). *Beyond Machiavelli: Policy analysis comes of age*. Washington, DC: Georgetown University Press.

Roberts, L. (2006, May 17). Students who fail AIMS test look for someone to blame. *The Arizona Republic*.

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Educational Policy Analysis Archives*, 11(24). Retrieved November 8, 2006, from http://epaa.asu.edu/epaa/v11n24/

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research and Evaluation*, 7(14). Retrieved September 18, 2008 from http://pareonline.net/getvn.asp?v=7&n=14

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it*. Cambridge, MA: Perseus Books.

Scheurich, J. J., Skrla, L., & Johnson, J. F. (2000). Thinking carefully about equity and accountability. *Phi Delta Kappan*, 82(4), 293–299.

Schrag, P. (2000, January 3). Too good to be true [Electronic version]. *The American Prospect*, 11(4). Retrieved September 18, 2008 from http://www.accessmylibrary.com/coms2/summary_0286-27373465_ITM

Schiller, K. S., & Muller, C. (2000). External examinations and accountability, educational expectations, and high school graduation. *American Journal of Education*, 108(2), 73–102.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practice*, 9(3), 15–22.

Smith, M. L. (1991). Meanings of test preparation. *Educational Research Journal*, 28(3), 521–542.

Stake, R. E. (2001). *Evaluation of testing and criterial thinking in education*. Paper presented at the annual meeting of the American Psychological Association, San Francisco, CA.

Stotsky, S. (1998). Analysis of the Texas reading tests, grades 4, 8, and 10, 1995–1998. *EducationNews.Org*. Retrieved July 8, 2002, from http://www.educationnews.org

Swope, K., & Miner, B. (Eds.). (2000). *Failing our kids: Why the testing craze won't fix our schools*. Milwaukee, WI: Rethinking Schools, Ltd.

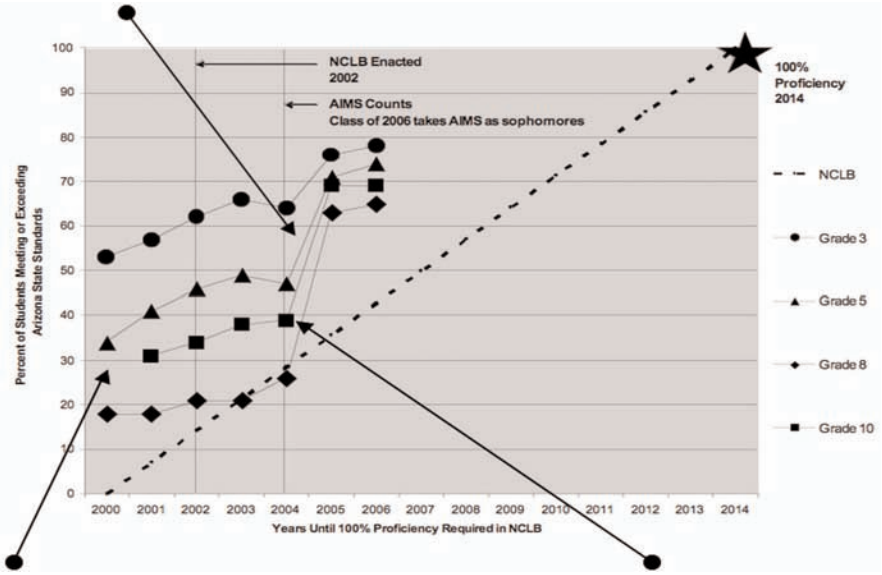
Urdan, T. C., & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy*, 8(2), 137–157.

Viadero, D. (2000). High-stakes tests lead debate at researchers' gathering. *Education Week*. Retrieved November 12, 2006, from http://www.edweek.org

Yara, G. (2006, April 5). Let AIMS wait begin. *The Arizona Republic*.

Figure 1.

The jump in scores from 2004 to 2005 occurred when the former AIMS tests were revised and the AIMS Dual Purpose Assessments (DPAs) were first administered. The revised tests were "the first to be aligned with the state standards" but were arguably easier than previous forms of the AIMS (see also Stotsky, 1998). New cut scores for the tests were also established (Arizona Department of Education, 2005).



The 2000 AIMS high school exit exam (grade 10) math scores are not listed because the assessment did not focus on core mathematics skills nor was it comparable to the content of the previous AIMS assessments (Arizona Department of Education Standards and Assessment Division, 2002).

The high school class of 2006 first took the AIMS as sophomores in 2004. Thirty-nine percent of the students who took the AIMS math test passed on their first attempt. In 2005, the test was changed. When the class of 2007 first took the AIMS as sophomores, 69 percent passed on their first attempt. Thirty percent more sophomores passed the AIMS math exit exam from 2004 to 2005.