A New Vision of Authentic Assessment
to Overcome the Flaws in High Stakes Testing

Dan French
Center for Collaborative Education, Boston, MA

In addition to test scores, parents and teachers want to see portfolios, exhibitions, performances, and demonstrations that show how well students achieve state standards.
A New Vision of Authentic Assessment

to Overcome the Flaws in High Stakes Testing

by Dan French

Recently, a former biologist, who is now a teacher intern in Alaska, wrote to his professor the following in an e-mail discussion about testing:

We inadvertently ran a test at the hatchery while I was there…. One culturist sampled his fish every week to check growth rates and calculate the feed, while the other culturist just fed what the fish would eat. Guess what? After six weeks the fish that were left undisturbed had almost twice the growth rate of the ones that were sampled constantly. Why? Stress levels [of the fish]. Our kids are no different. Stress was incredibly high during testing week at my high school…and noticeable for a while after in some of my classes. (Jones, 2002)

The advent of the No Child Left Behind federal legislation has plunged us into an unprecedented era of high stakes testing, with the presumption that testing and more testing, coupled with the threat of not being promoted from grade to grade and of not graduating from high school, will be the engine that drives improvement in instruction and student achievement. Yet there are many inherent flaws to this approach, an approach that, under a mantel of equity and excellence, threatens to undermine the tenets of exemplary middle grades practice, and leave behind the very students that the legislation and testing movement purport to be helping.

The Two Faces of Standards in Addressing Educational Equity

In a society like ours that is stratified by race and income, we must have standards of what all students should know and be able to do upon graduating from middle to high school, and from high school. The absence of standards virtually guarantees stratified resources and access to knowledge, based upon income, color of skin, and the community and neighborhood in which one lives. For example, two Massachusetts Department of Education studies, prior to the state enacting education reform legislation in 1993, found profound differences in the coursework of algebra and U.S. history courses in suburban, rural, and urban districts – suburban students were given more homework, had longer class periods, had greater access to educational materials, and received a more rigorous curriculum than urban and rural students enrolled in courses with the same title (Quincy: Massachusetts, 1986, 1985). One study reported that, “The conclusion from all the data available is that the curriculum taken by minority students is not the same as that taken by white students.” (Mass. DOE, 1986, p.7) Standards form a tool to promote high quality curriculum and instruction for all.

Standards that promote equity are broad and streamlined in nature; they are designed to ensure that every student graduates with the habits of mind, skills, and knowledge to be a contributing member of a democratic society and world and to be prepared for a future productive life. Broad, streamlined middle grades standards promote relevant and rigorous instruction and curriculum that builds upon the principles of early adolescent development. These standards allow for locally determined assessments that require students to demonstrate mastery over essential knowledge through varied means, including portfolios, demonstrations, and exhibitions.
However, standards can also reinforce and amplify the current inequities that pervade public education. It is with this understanding that the National Forum to Accelerate Middle-Grades Reform recently released the following statement on high stakes testing:

The National Forum believes in standards and assessments that lead to high expectations, foster high quality instruction, and support higher levels of learning for every student. At the same time, the National Forum believes that no single test should ever be the sole determinant of a young adolescent’s academic future, whether it be promotion to the next grade, special placement, or transition from the middle grades to high school. (2002)

With this lens, we need to take a careful look at the impact of linking state-level high stakes testing to the standards movement, and the resulting impact on curriculum and instruction. Ultimately, we need to examine its effects on student achievement, particularly those students that federal and state governments claim the standards and testing movement is most designed to help—low-income students and Black and Latino students. We need to examine the industry claims of validity and reliability of high stakes standardized tests, and question the very reasons our government and the education industry have embraced these tests. Finally, we need to ask ourselves if there is another way to employ standards as an equity tool that leads to more effective middle grades practice and that better assists every student to achieve at high levels.

**No One Measure Can Do Justice in Assessing a Student’s Knowledge**

Rothstein (2000) used a baseball example to question the veracity of using a single high stakes test to measure a student’s knowledge:

Mike Piazza, batting .332, could win this year’s Most Valuable Player award. He has been good every year, with a .330 career batting average,…and a member of each All Star team since his rookie season. The Mets reward Piazza for this high achievement, at the rate of $13 million a year.

But what if the team decided to pay him based not on overall performance but on how he hit during one arbitrarily chosen week? How well do one week’s at-bats describe the ability of a true .330 hitter?

Not very. Last week Piazza batted only .200. But in the second week of August he batted .538. If you picked a random week this season, you would have only a 7-in-10 chance of choosing one in which he hit .250 or higher. (p. B11)

Rothstein questioned the validity of assessing a student’s knowledge at one point in time. Studies have documented that how students fare on standardized tests can be greatly influenced by a host of external factors, including stress over taking the test, amount of sleep, distractions at the testing site, time of day, emotional state, and others (Kamin, 1974; Sacks, 2000). Indeed, these influences most dramatically affect low-income students and students of color. Steele has found that African American students perform significantly below their potential on tests in which they know they are being assessed on their intellectual performance, largely due to what he calls “stereotype threats,” a historical and well supported fear that tests created by the dominant culture will judge them to be wanting (1999). We only need revisit the origins of testing in the United States – that of the eugenics movement and the rise in immigration in the early 1900s, resulting in IQ and Stanford-Binet tests as a means to sort and track students primarily.
based on race and income (Sacks, 2000)–to understand that the roots of high stakes testing lie in sorting students.

Standardized tests are also poor predictors of how well students can apply the knowledge that they do demonstrate on these tests. A recent study examined data from 18 states that have implemented high stakes testing programs to assess whether students gained any knowledge that they could apply elsewhere, other than learning the necessary facts for performing on a state’s high stakes test. (Amrein and Berliner, 2002). They concluded that, “Analyses of these data reveal that if the intended goal of high stakes testing policy is to increase student learning, then that policy is not working. While a state’s high stakes test may show increased scores, there is little support in these data that such increases are anything but the result of test preparation and/or the exclusion of students from the testing process” (Amrein and Berliner, p. 2).

Kane and Staiger’s 2001 research study found that “between 50 percent and 80 percent of the improvement in a school’s average test scores from one year to the next was temporary and was caused by fluctuations that had nothing to do with long-term changes in learning or productivity” (as cited in Olson, 2001, p.9). According to Kane and Staiger, these findings are caused by the variations in student population in a given grade from year to year, variations in test conditions, and the presence of rewards and punishments from the state or district. Other studies have concluded that students “lost” from the testing pool – through grade retention, dropping out, or being excluded from taking the test because of having special education or bilingual status – create the appearance of an improving school when in fact the opposite may very well be true (Test scores unreliable means of assessing school quality, 2001).

Most importantly, standardized test scores are far from an accurate predictor of how students will do in later life, even as they take on the draconian role of determining a student’s future life opportunities in many of our states. As former Harvard and Princeton presidents Derek Bok and William Bowen attest in their book The Shape of the River:

[They] researched what became of 700 African-American students who would have been bumped out of their chosen schools [based solely] on test scores. About 70 of the freshmen became doctors, about 60 became lawyers, about 125 became business executives, more than 225 earned professional degrees or doctorates, more than 300 are leaders of civic activities. The average [annual] earnings of the 700 are more than $71,000. “Taken together, grades and scores predict only 15 to 20 percent of the variance among all students in academic performance and a smaller percentage among black students.” (as cited in Jackson, 1999, p.A23)

Are These Tests Even Reliable?

The high error rate that has been experienced in virtually every major testing program and state in the nation in which high stakes tests are used calls to question the validity of such tests. The number of test errors by the major testing companies are too numerous to detail, with many having disastrous effects on the students they are measuring. Some highlights include:

1. A Massachusetts history teacher found an error in one multiple choice question on the state’s 8th grade high stakes test which resulted in 666 students failing the test who would have otherwise passed, many of whom were retained in grade as a result (Vaishnav, 2002).
2. In Minnesota in 2000, the testing company, NCS Pearson, made a scoring error on the tests of 47,000 students; of these, almost 8,000 eighth through twelfth grade students incorrectly received failing grades, with 54 high school seniors being denied diplomas until the scoring error was discovered (Henriques and Steinberg, 2001).

3. In New York City in 1999 nearly 9,000 middle and high school students were mistakenly required to attend summer school because of test scoring errors by CTB/McGraw-Hill. McGraw-Hill admitted that they also made scoring errors in Indiana, Nevada, Tennessee, South Carolina, and Wisconsin (Henriques and Steinberg).

4. In November 1998, Harcourt Brace misclassified as “below basic” or “basic” elementary students who had actually scored at “proficient” or “advanced” levels. The mistake was detected by a principal (Error affects test results in six states, 1999).

High error rates and poor reliability are some of the reasons why a multitude of respected national education organizations, including the National Forum for Accelerating Middle-Grades Reform and the American Educational Research Association, have issued public statements decrying the use of a single high stakes test to determine students’ futures.

The Hidden Consequences on Students of High Stakes Testing

While proponents of high stakes tests declare that their approach has led to increased gains in student achievement, and even narrowed the achievement gap by race and income, this is hardly the case. On the contrary, the students who are most negatively affected by the consequences of high stakes tests are low-income, Black and Latino (McNeil, 2000; Orfield and Wald, 2000). Massachusetts and Texas, leaders in the high stakes testing movement and widely ballyhooed as exemplary systems, are good cases to examine.

In Massachusetts, since the advent of their high stakes test, the Massachusetts Comprehensive Assessment System (MCAS), the percent of middle grades dropouts has increased. These dropouts are increasingly African American and Latino, 19 times and 14 times, respectively, more likely to drop out than their White peers. These students, who must be 16 in order to legally drop out, are the likely products of multiple years of being retained in grade. Grade retention, a critical correlate to dropping out of school (Smith and Shepard, 1989), is on the rise and at all-time highs in many urban districts, as these districts strive to prevent students from entering grades with high stakes tests until they are deemed ready to pass (Clarke, Haney, and Madaus, 2000). The number of students who are unaccounted for within each year’s cohort of students as they pass from grade to grade has doubled. Most of these students have probably been held back in grade (Wheelock and Haney, 2002).

Boston College’s Progress Through The Educational Pipeline Project indicates that in this trend, nearly one out of three members of the Massachusetts class of 2003 that began in the ninth grade may not be graduating in June 2003. The numbers are more disturbing when disaggregated by race. Less than half of all Black and Latino students who were in eighth grade in 1998 and are members of the graduating class of 2003 have passed the
MCAS as of spring 2002. Latino students are four times as likely, and Black students are three times as likely to fail the MCAS as are White students. Low-income students are three times as likely to fail the MCAS as are their more affluent peers (Wheelock and Haney, 2002). Disturbingly, a study has found that the MCAS reflects an increased disparity in achievement by race when compared to student’s performance in classrooms. The achievement gap between White students and African-American and Latino students has been found to be significantly larger on the MCAS than it is for students’ classroom math grades (Brennan, Kim, Wenz-Gross, and Siperstein, 2001).

Similar findings have occurred in Texas, the birthplace of the No Child Left Behind Act. In a study on the effects of the Texas Assessment of Academic Skills (TAAS) on Black and Latino students, twice as many Black and Latino students were found to have failed this high stakes test as had White students (Orfield and Wald, 2000; McNeil, 2000). Simultaneously, Texas’ dropout rate has dramatically increased as has the percent of students being exempted from taking the TAAS because of special education and bilingual placement, while the college attendance rate remains among the lowest in the nation. In all these cases, students are disproportionately Black and Latino. The study’s authors concluded that high stakes standardized tests will continue to hinder the ability of African American and Latino students to graduate from high school (Orfield and Wald, 2000; Haney, 1999; McNeil, 2000).

Narrowing the Window on Instruction, Curriculum, and Assessment

This maze of high stakes tests invariably has its impact on where it matters most – the classroom and the relationship between teachers and students. In many states in the middle grades, we have witnessed a narrowing of the curriculum and a movement away from the type of curriculum and instruction that best assists middle grades students to learn and become productive citizens in our larger world. Studies have found that high stakes testing decreases choice in the classroom, decreases motivation to learn, focuses the curriculum on coverage at the expense of depth and understanding, and narrows the scope of instruction to more teacher-centered strategies rather than student-engaged instruction (Sacks, 2000).

In fact the high stakes testing movement impedes the drive toward creating more academically challenging courses of study for middle grades students and inhibits the creation of collaborative faculty teams of teachers engaged in discourse on improving instruction. In Massachusetts, which administers a high stakes test, all schools are required to develop individual student plans for any student that fails the math or English portion of the exam. Unfortunately, in some schools that have high percentages of low-income students, students of color, and students whose first language is other than English, upwards of 80% of students fall into this category. The result is that tremendous amounts of team common planning time are spent writing individual student plans, that end up looking the same, taking away valuable time from teachers’ discussions of how to improve instruction through practices such as looking at student work, peer observation, data-based inquiry, and action research (R.Tung, personal interview, May 18, 2002).

Other examples in Massachusetts’ middle schools include the following:

1. In one district, in response to low district-wide math scores, the superintendent required an additional 300 minutes of math instruction per
In one middle school in which the arts is their magnet theme, the only place in the schedule where there was flexibility to give up time was in the arts. Thus, in a school in which the arts is a tremendous drawing card to motivate low-income and Latino students to attend school, the arts has been severely curtailed. Further, as there were no additional math teachers in the district, teachers who are uncertified in math are now teaching the extra math.

2. In one school in another district, an acclaimed program in which eighth grade students spend a week in the spring apprenticing with various city businesses, nonprofits, and city agencies, has been curtailed because of the need to devote more time to test preparation.

3. In multiple middle schools, anywhere from two to four weeks before the tests begin, all classes stop and test-preparation begins, with students taking practice test problems day after day. Adding in the actual testing period, a significant percentage of students’ school years are sacrificed to testing (McNeil, 2000).

Ultimately, we are creating an education system in which depersonalization in teaching and learning is at the core. We should take to heart the following quote by Annelise Schantz, the valedictorian of the 2000 graduating class at Hudson High School in Massachusetts:

So I’m the valedictorian. Number one. But, what separates me from number two, three, four, five, six, 50, or 120? Nothing but meaningless numbers. All these randomly assigned numbers reflect nothing about the true character of an individual. They say nothing…about desire or will. Nothing about values or morals. Nothing about intelligence. Nothing about creativity. Nothing about heart. Numbers cannot and will not ever be able to tell you who a person really is. Yet in today’s society we are sadly becoming more and more number oriented. Schools today are being forced to teach to the numbers…. The MCAS serves as just another set of meaningless numbers that add one more reason to focus on scores and forget learning…. Judging us by our competency on a biased test is perhaps the biggest injustice that the state could ever inflict upon us…. Does anyone care about the human beings behind the numbers? (Schantz, 2000)

Is This What the Public Wants?

If high stakes testing is having deleterious effects on the middle grades, why are we so passionately rushing down this path? Whose agenda is it? Do we really believe that highly prescriptive standards and high stakes testing present the ultimate solution to the complex and messy challenge of educating our increasingly diverse early adolescents?

This agenda is certainly not being driven by teachers and parents, those who best know what it takes to effectively educate our younger generation. An August 2000 Gallup Poll produced for * Phi Delta Kappan* found that only 13% of the public believed that a single standardized test was the best way to measure academic achievement, while 85% felt that the best way was either a combination of standardized and teacher-designed tests or portfolios of students’ work and other demonstrations of academic competence. In a 2001 Public Agenda poll, 75% of parents and 90% of teachers felt that it is “wrong to use the results of just one test to decide whether a student gets promoted or graduates.” In this same poll, 83% of teachers feared that “teachers will end up teaching to the tests instead of making sure real learning takes place,” and 82% felt that “schools today place far too much emphasis on standardized tests” (*Education Week*, pp. S4,7).
What we have, then, is a high stakes testing movement being fueled, not by those who best know and care about the students in our middle schools, but by others outside of our public schools who have varied interests at heart. In 1999, NCS Pearson, a large testing company, reported more than $620 million in revenues, up 30% from the previous year. McGraw-Hill, another large testing company, also owns programs such as Open Court and Reading Mastery, two direct instruction programs that are being purchased in large numbers by districts striving to drive up their standardized test scores (Kohn, 2002). State spending on testing has increased multifold. It is estimated that the K-12 standardized testing industry is as much as $1.5 billion per year (Kohn, 2002; Gluckman, 2002). Business leaders and legislators have lined up behind this industry as a quick fix to the dilemmas of educating a diverse student population.

**Summing Up the Damage**

Excellent middle schools have as their foundation the following four strengths: rigor, relevance, relationships, and resources:

1. Rigor, or academic challenge which is personalized, that builds upon the relationships and experiences of young adolescents, and that prepares students to be literate, participating members of a democratic society;
2. Relevance, in that the curriculum, instruction, and assessment has meaning, is developmentally responsive, and makes connections to students’ lives and the world around them;
3. Relationships, in which middle schools are small and personalized enough so that every student is known well by at least one caring, nurturing adult in the school, and in which positive student relationships are fostered and valued within the classroom (in fact, Bryk and Schneider suggest a correlation between strong, positive relationships and high trust levels with students’ academic improvement (2002)); and
4. Resources, in which middle schools have maximum, charter-like control over budget, staffing, curriculum, governance, and time to best meet the needs of their unique student population, and, in the words of the National Forum to Accelerate Middle-Grades Reform, are used “to work to overcome systemic variation in resources and outcomes related to race, class, gender, and [achievement]” (National Forum to Accelerate Middle Grades Reform, 1999, www.mgforum.org/vision.asp).

Unfortunately, the high stakes testing movement has latched onto only one of these four criteria—rigor. Absent the other three criteria for excellent middle schools, high stakes testing ends up adopting the very narrow definition of rigor outlined in The American Heritage Dictionary: “Strictness or severity…; a harsh or trying circumstance…; a harsh or cruel act” (1993). Such conditions inevitably end up harming those students and schools who are most vulnerable.

Ultimately, state-level high stakes testing is a fundamentally flawed concept, and does not achieve the stated ends of equity and excellence. The current standardized testing movement undermines the vision of creating excellent and equitable middle grades schools that are academically challenging, developmentally responsive, and socially equitable by creating a narrow and rigid view of learning and teaching that
exacerbates the inequities of our educational system and larger society. A single score on a test should never stand as the sole measure of a student’s knowledge, understandings, performance, and intellectual habits, and should never be used to determine a student’s future.

**Are There Alternatives? Crafting a New Vision of Assessment**

Are there alternatives to the high stakes testing movement that can be a catalyst for creating high performing middle schools, in which equity is as much of a banner as is excellence? In its policy statement on high stakes testing, the National Forum to Accelerate Middle-Grades Reform “encourages diverse approaches to curriculum and instruction and supports the use of multiple measures to make decisions about a student’s progress. These may include portfolios, exhibitions, performances, demonstrations, and tests that measure how well students achieve state standards” (National Forum to Accelerate Middle Grades Reform, 2002). This statement acknowledges that standardized tests do have a legitimate role in assessment, mostly as a diagnostic tool, yet makes clear that other assessment forms are necessary. Do these types of authentic assessments have value? Is another means of testing other than state-level high stakes standardized tests possible, and if so, do they improve instruction and student achievement?

While the research is quite preliminary, there are some promising findings as to the impact of authentic assessment practices on student learning. While standardized tests have their limitations in making high stakes decisions, they do provide one legitimate, albeit narrow, form of tracking student progress. The National Center for Research on Evaluation, Standards, and Student Testing’s study on the effects of a year-long focus on classroom performance assessment found small but real academic gains, including gains on low-stakes standardized tests in mathematics. The Center concluded that students had “opportunities to develop their mathematical understandings that had not occurred previously” and that did not occur with the control group of classrooms which did not engage in performance assessments (Shepard, 1995, p.14). A study of the Chicago Annenberg Research Project found that when teachers used high quality assignments that contained higher order thinking, in-depth understanding, elaborated communication, connections to students’ lives beyond school, students produced higher quality work and even achieved greater than average gains on low-stakes standardized tests in reading, writing, and mathematics (Newman and Nagoaka, 2001).

What would an assessment system look like that promoted the principles of rigor, relevance, relationships, and resources? First, authentic assessment needs to start at the school level, framed by a philosophy of assessment that contributes to excellent teaching and increased student learning. The Boston Pilot Schools Network, a group of 13 schools, five of which contain middle grades, and are members of the Boston Public Schools yet have charter-like autonomy over their resources, have articulated a helpful framework for thinking about assessment and accountability:

The Pilot Schools believe that a strong system of assessing student progress is vital to creating excellent schools in which all students learn and achieve at high levels. We believe in standards that lead to excellent schools, not standardization…. Ultimately, good assessment systems should open doors for all students rather than shut them, and help students graduate with a range of options.
The network’s assessment system is built upon the following principles:

- Provide multiple ways of assessing student competency in meaningful ways
- Eliminate secrecy so that all students, families, and the public understand what students should know and be able to do
- Develop the assessment system by those working most closely with students
- Embed good assessment in curriculum and instruction that engages students in work that has a public purpose
- Help students become independent, self-reliant, and thoughtful learners through good assessment
- Promote reflective practice in teachers through good assessment (Pilot Schools Network, 2000).

*Turning Points*, a New American Schools middle school reform model based on the seven guiding principles of effective middle grades reform framed in the *Turning Points 2000* report, articulates a similar philosophy of assessment. The *Turning Points* design promotes assessments in which students are required to demonstrate that they can do important work, and in which there are essential knowledge, skills, and habits of mind that students know and are able to articulate in multiple ways. In *Turning Points* schools, teachers devise a wide array of assessments—including portfolios, exhibitions, theses, and demonstrations—in which students are required to exhibit understanding, analysis, synthesis, evaluation, and application. For example, eighth grade exhibition and portfolio exit reviews, fortified by rubrics and panels of reviewers, replace graduation by course counting and seat time. In this example, high stakes at the state level is replaced by high stakes in schools and classrooms, a more rich, relevant, and meaningful process that is embedded in the daily curriculum and lives of students.

Can such a process of rich, school-based authentic assessments bring instructional improvement and raised student achievement to our middle schools? How do we measure achievement when schools are using different forms of assessment? The Massachusetts Coalition for Authentic Reform in Education, a volunteer group of parents and educators who seek to influence the state to adopt a more comprehensive approach to assessment than merely state-level high stakes testing, offers a compelling vision of state-level accountability:

1. Streamlined competencies, rather than the reams of endless, mind-numbing standards that predominate in most states, which can be given to entering middle grades students and their parents in two to four pages, that outline the broad habits of mind, skills, and concepts that students should know and be able to do at the end of eighth grade;
2. Standardized testing solely in literacy and numeracy to provide one legitimate method of publicly tracking the progress of schools and students from year to year, both their overall progress and their progress in closing the achievement gap between low-income students and their more affluent peers, and between White students and Black and Latino students;
3. Locally developed and administered authentic assessment systems—focused on exhibitions, writing across a range of genres, portfolios, and
demonstrations, and approved by the state—that measure whether students have mastered the state competencies;

4. A school quality review model, in which external teams of practitioners assess the effectiveness of a school’s practices in delivering high quality and equitable instruction to all students, based on the successful models in Britain, Rhode Island, and Massachusetts’ charter schools; and

5. Required annual local reporting by schools to their communities, using a defined set of multiple indicators—including dropout, suspension, attendance, grade retention, and course failure rates, enrollment in high level classes, and achievement on both standardized tests and authentic assessments—with a focus on both student outcomes and on equal opportunity and access to knowledge for all students.

Such a system, while admittedly more complex than state-level high stakes tests, will give a truer picture of schools and their students, one that enriches instruction, informs parents, students, and teachers, and promotes reflection on how to improve. It can be accomplished.

Conclusion

The work of creating academically challenging, developmentally responsive, and socially equitable middle schools that serve a diverse range of students is much more complex and messy than merely measuring students using a high stakes standardized test. We need to acknowledge the systemic reasons why low-income students and Black and Latino students are achieving at far lower rates than their White, more affluent peers and are much less engaged in the institution called school, and identify the conditions that create excellent schools for diverse students. Simple solutions to complex problems usually just create more problems. Let’s get back to the real business of figuring out how to create excellent middle schools for each and every child.

Dan French is the executive director of the Center for Collaborative Education, Boston, Massachusetts.
E-mail: dfrench@ccebos.org
References


Quincy: Massachusetts Department of Education. (1985, June). Course taking among Massachusetts high school students.

Quincy: Massachusetts Department of Education. (1986, April). The high school experience in Massachusetts.


