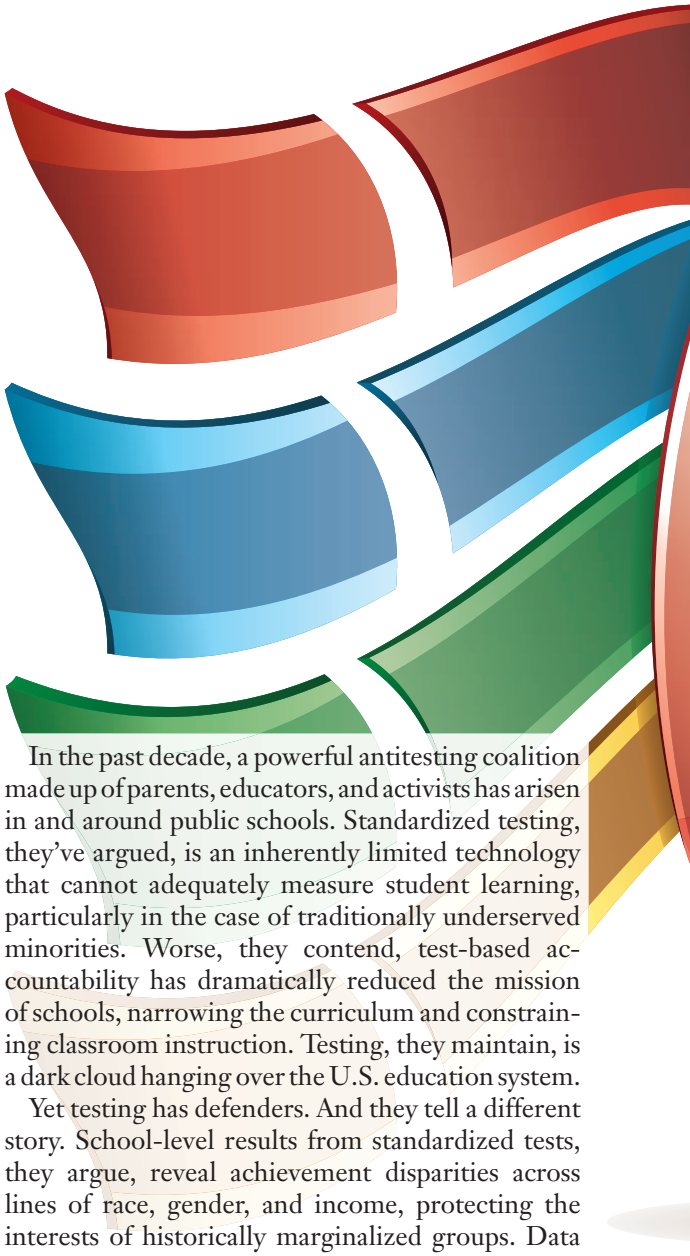
# The best of both worlds

Replacing machine-scored standardized tests with teacher-rated classroom assignments and accurate grading may represent our best hope for promoting both accountability and instruction.

**By Jack Schneider, Joe Feldman, and Dan French**



In the past decade, a powerful antitesting coalition made up of parents, educators, and activists has arisen in and around public schools. Standardized testing, they've argued, is an inherently limited technology that cannot adequately measure student learning, particularly in the case of traditionally underserved minorities. Worse, they contend, test-based accountability has dramatically reduced the mission of schools, narrowing the curriculum and constraining classroom instruction. Testing, they maintain, is a dark cloud hanging over the U.S. education system.

Yet testing has defenders. And they tell a different story. School-level results from standardized tests, they argue, reveal achievement disparities across lines of race, gender, and income, protecting the interests of historically marginalized groups. Data from standardized tests have spurred schools and districts to intensify their focus on student achievement, fostered coordination among state and district offices, and provided objective information that enables the public to hold schools accountable. Achievement data, they argue, function as a form of sunlight.

Insofar as these sides are diametrically opposed to each other, it's impossible to have a foot in each

**JACK SCHNEIDER** (jschneid@holycrossedu, @edu_historian) is an assistant professor of education at the College of the Holy Cross, Worcester, Mass., and director of research for the Massachusetts Consortium for Innovative Education Assessment (MCIEA). **JOE FELDMAN** (joe@crescendoedgroup.org) is CEO of Crescendo Education Group, Oakland, Calif., and is a former high school teacher, principal, and district administrator. **DAN FRENCH** (dfrench@ccebos.org) is executive director of the Center for Collaborative Education, Boston, Mass., and a co-founder of the MCIEA.

**Classroom-based data are superior to standardized test data if teachers validly measure student performance and develop ways of reliably reporting on it.**

Thinkstock/iStock

camp. One must choose either to support or to reject data — opting in or opting out — despite the fact that each side makes valid points.

But this needn't be the case.

The opportunity to find common ground starts by shifting the debate away from whether to collect data and focusing, instead, on what data to collect. After all, standardized test scores are not the only information we have about student achievement. In fact, they aren't even the most common. Classroom assessments, evaluated by teachers, are given both more frequently and more broadly than standardized tests. And they're authentically embedded into classroom instruction. Insofar as that's the case, such assessments might offer a more comprehensive picture of student learning and have a less distorting effect on curriculum and instruction.

The problem, of course, is that evaluation of student work — most commonly reported in the form of A to F grades — is not consistent across settings. A B issued in one school, for instance, often doesn't represent the same level of proficiency as a B at another school. Just as frequently, the meaning of a grade can differ from teacher to teacher within the same school, and even from student to student within a single classroom. Such variance seemingly makes teacher-assessed assignments inappropriate for the purpose of external measurement and accountability.

Yet what if that were to change? What if we could rely on teachers' assessments for the information currently provided by standardized test scores? Using student work in this manner would save instructional time, better capture the true abilities of diverse students, and reduce the problem of teaching to the test. In addition, it might resolve the ongoing feud over data that distracts from the actual work of improving schools.

This is not to say that such work would be easy. Replacing machine-scored standardized tests with teacher-rated classroom assignments would require the investment of time and the outlay of resources. And it would require serious district, state, and national commitment. Yet such work may represent our best hope for promoting not only both accountability and instruction — but also a system that captures useful information while strengthening learning.

## A history lesson

Surprising though it may seem, graded student work has not always been a feature of K-12 schools. In the small and informal school settings of the colonial and early republic periods, teachers knew how students were performing and could communicate with parents and other educators through face-to-face conversations. The closest thing to a "grade" was an end-of-year student performance that often took place in a public setting — a culminating exercise designed to shore up the community's trust that learning was taking place.

In the second half of the 19th century, however, as schools grew larger and became more systematized, grades emerged as a communication device, conveying information from teacher to teacher, school to school, and school to family. Still, grades were never used to communicate beyond the school community. In fact, until the early 20th century, that would have been impossible. Teacher rating of student work took various forms — narratives, 1-10 scales, A-Z systems, and so on — with no standard across communities. Instead, parents and educators within a school simply calibrated themselves to a set of particular norms. By the first decades of the 20th century, a standard model — the A-F system — had emerged. But for district leaders and state policy makers, grades still didn't represent a uniform enough currency for measuring schools. Without a common definition about what grades meant — in other words, without cross-school validity and reliability — they turned to tests.

The first standardized tests were actually created several generations earlier, in Boston during the 1850s. Civic leaders developed the tests for the general purpose of measuring school quality across the city and for the specific purpose of exercising control over increasingly powerful and autonomous school principals. In other cities and states, policy elites with similar interests developed their own tests. The first statewide standardized testing, for instance, emerged in New York in the wake of the Civil War. On the whole, these tests were similar in form. Like modern standardized tests, they focused primarily on factual knowledge for ease of scoring. They were mass-produced to keep costs down. And by the early 20th century, they increasingly used a new format: the multiple-choice question. Tests weren't yet scored by machines, and school-level scores were often simply compared to "expected average" scores provided by the test producer. But much would be recognizable to modern observers, including the inordinate time devoted to testing. As the president of the New York teachers union put it in 1930, the state Regents examination was a "continuing waste of childhood that is appalling to contemplate."

Despite complaints — about loss of instructional time to testing, inaccuracy of the measures, cost, and infringements on teacher autonomy — standardized testing continued to expand across the 20th century. In a publicly funded and decentralized system, tests offered a mechanism for accountability and governance. And, in an ostensibly meritocratic system that determines many life outcomes, tests offered a seemingly fair way of determining social mobility. Consequently, the public generally supported testing, as

did an emerging web of testing experts, corporate developers, and state accountability systems. Soon enough, testing simply became a part of standard operating procedure in schools; it became a cultural norm.

Historically, then, we have two separate and highly idiosyncratic systems for doing similar kinds of work. Within the school community, teacher assessment of student learning, most commonly in the form of grading, is a well-accepted practice for measurement and communication. Outside of the school, however, standardized tests are the accepted mechanism for evaluating student achievement. Thus, although most teachers would scoff at the idea of using standardized tests to measure student learning within their own classrooms, the practice is dominant when it comes to communicating about achievement beyond the school community.

**We must begin by shifting the conversation away from a false choice – between data and no data – to one about how to make data useful.**

Each system has continued on, not because it represents the best we can do, but because it's all that most of us have ever known. As challenges have arisen, though, they have spurred questions. What if things could be different? What if it were possible to keep the strengths of each system, preserving the curricular relevance and informational richness of teacher-led assessment, while also maintaining measures of reliability and validity currently associated with standardized tests?

## Promising practices

Historically separate though they may be, these parallel and competing systems for measuring student learning are not incompatible. As unlikely as it sounds, modern examples at the local and state levels demonstrate different approaches to ensuring that teacher assessment of student work has the measurement properties previously attributed only to standardized tests.

### *A focus on reporting: Standards-based grading*

Take the example of San Leandro High School in Northern California (SLHS), where one of us has been working over the past year to ensure that teacher-issued grades function as reliable and valid descriptors of student academic performance. SLHS is the only comprehensive high school in the San Leandro district, serving 2,600 students in grades 9-12. The school is quite diverse: 45% Hispanic/Latino, 18% African-American, 15% Asian, 9% Filipino, and 9% white, with 13% of the students classified as English learners, 12% special education, and 66% qualifying for free or reduced-price lunch. The district administration, having undertaken California's adoption of the Common Core standards and its own initiative to implement project-based learning, realized that evaluation and grading needed to be examined as well. As the district's director of teaching, learning, and equity said, "If we're going to ask teachers to unpack new standards and learn new instructional approaches, then we need to unpack the systems that have been barnacled in our schools, like grading."

Through a bottom-up process, this work, sometimes called "standards-based grading," focused primarily on building capacity among the high school's faculty. As such, a subset of teachers, which included all department chairs, began with a study of research-based practices that can improve grading — practices like using rubrics to clearly describe student performance levels, employing a 0-4 point scale rather than a 0-100 percentage, assigning grades based on a student's most recent performance rather than on his or her work to date, and limiting the weight of formative assessment data like homework. Afterward, a teacher group piloted these practices, which they collected data on and refined through a series of action research cycles.

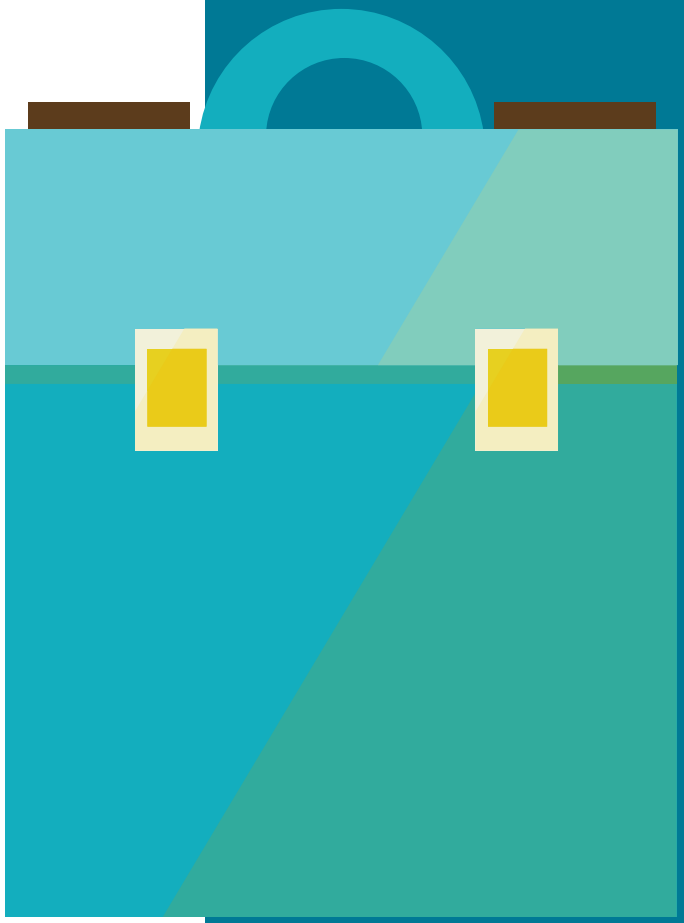
Lucy, a 10th-grade English teacher and chair of the school's English department, provides an example of a teacher's experience in this pilot group. Her grading had been essentially unchanged over her 18-year career, consistent throughout every instructional change and textbook adoption, and she was skeptical of the need and the value of examining grading. Still, she was open to new ideas. Throughout the year, she prototyped new approaches to evaluating and grading. Some of the changes were technical — increasing the percentage weight of academic performance in the grade and reducing the weight of nonacademic behaviors — while others were more substantive, such as allowing retakes, using checklists and rubrics, and not including formative assessment results in a grade. Lucy found that her new grading practices improved student learning and more accurately reflected student proficiency on the standards. Compared to the previous year, the percentage of D's and F's she awarded decreased, as did the percentage of A's.

Across the pilot cohort, grading became less idiosyncratic and subjective, and more descriptive of

***Join the conversation***

*facebook.com/pdkintl*
*@pdkintl*

**Portfolios of student work evaluated and reported consistently across teachers can be credible enough to render standardized tests unnecessary.**



standards mastery, supporting every student's success. As Maritza, the co-chair of the world languages department, explained, "Giving retakes and replacing a student's previous score with the improved score doesn't mean that I'm lowering my standards; it means that I am preparing the students for real life. I give them as many opportunities as I can to show me that they have learned. I am helping the students to see they are smart and they are capable; they need to study, and even if they need more time, they are able to reach their goals just as well as the other students."

In response to the pilot group's results and its strong endorsement of the experience, the district has decided to expand the examination of grading. Over the next two years, the district has committed to have every teacher in both middle schools and the high school — over 200 — participate in this process of action research cycles to improve grading. Ultimately, teachers will amass a body of evidence and experiences so they can develop common and research-based grading practices within and across grade levels, departments, and schools. This effort is intended to inform the district's other instructional initiatives, ultimately creating consistent expectations of standards performance levels with a grading and reporting system that reliably and accurately reports that performance.

### *A focus on measurement: Performance tasks*

Although San Leandro chose to improve grading and then develop common expectations of student performance, efforts to improve teacher assessment of student work can happen in the reverse order as well. For the past two years, the New Hampshire Department of Education (NHDOE), partnering with several organizations including one of ours — the Center for Collaborative Education, has pursued what has grown to be an eight-district pilot project — Performance Assessment for Competency Education (PACE) — designed to anchor accountability in teacher-generated performance tasks rather than standardized tests. The federal Department of Education granted the state a waiver from No Child Left Behind to implement PACE as a pilot assessment and accountability system for a limited number of school districts.

To lay the groundwork for PACE, NHDOE facilitated practitioner committees to develop state-approved competencies in English language arts, mathematics, science, and the arts, aligned to the Common Core State Standards and Next Generation Science Standards. Unlike standards, competencies are broad learning targets representing key concepts and skills applied within or across content domains. For example, New Hampshire's high school Reading



Thinkstock/AbleStock

Literature Competency is: *Students will demonstrate the ability to comprehend, analyze, and critique a variety of increasingly complex print and nonprint literary texts*, while the grades 3-4 math numbers and number systems competency is: *Students will demonstrate an understanding of number systems, thinking flexibly, and attending to precision and reasonableness when solving problems using whole numbers, fractions, and decimals.* In addition, the state adopted a set of dispositions named Work-Study Practices, developed by a statewide committee of practitioners, that were deemed essential to student success — Creativity, Communication, Collaboration, and Self-Direction.

**Each system has continued on, not because it represents the best we can do, but because it's all that most of us have ever known.**

Next, cross-district teacher teams created common, competency-aligned, curriculum-embedded performance tasks for grades 3-8, as well as for grade 10 (which have yet to be released to the public). These common tasks were examined at the state level by staff at one of the partner organizations, National Center for Improvement of Educational Assessments, with reviewers providing feedback to teacher teams. After students completed the tasks, teachers individually scored their students' work, then cross-district teams assessed student work samples, collaborating to calibrate their scoring to ensure that teachers across districts scored student work at the same level of proficiency on four-point rubrics. Districts were then responsible for adding their own local performance tasks to supplement the common tasks in order to make determinations of student proficiency. For example, in Sanborn, a PACE district, three 4th-grade teachers embedded a local performance task within an interdisciplinary unit that integrates local civics standards and ELA Common Core State Standards in which students learn about government through persuasive writing. In pairs, students research a societal issue important to them. Each student then writes a persuasive essay, detailing arguments and counter-arguments, proposing legislation that would proactively address their chosen issue. The students then go back to their pairs to draft a final proposed bill, combining their ideas. The proposed bills are brought to a mock New Hampshire Senate where students present their bill to the Senate (their fellow students). With this task, students are introduced to the lawmaking process, the balances of power in government, and making decisions about civic issues that are important to them. As one teacher noted, "The persuasive writing [students] did was top-notch . . . because they cared about the issues they chose, and this choice is really important for engagement with the task."

Scores from common performance tasks have provided the state with comparability data to ensure cross-district reliability, whereas scores from local tasks, along with the one common task per grade per discipline, comprise the body of work (or evidence) used to make local determinations of student proficiency in each of English language arts, math, and science. Teachers determine what is included in the body of work for each discipline, such as work from local performance tasks and even locally administered standardized tests. They then use their best judgment to determine a proficiency rating for the student within that discipline, using each discipline's teacher-generated, state-approved Achievement Level Descriptors, which describe the knowledge and skills aligned with a given achievement level. Once again, cross-district sessions among teachers help establish reliability in scoring body-of-work ratings. Each district then reports to the state on student proficiency ratings, using these bodies-of-work ratings.

On the whole, these examples from both the school and state levels show that teachers can assess and report on student work in a manner that produces calibrated, reliable, and valid measures.

In other words, portfolios of student work evaluated and reported consistently across teachers can be credible enough to render standardized tests unnecessary.

**Join the conversation**

*facebook.com/pdkintl*
*@pdkintl*

## Some caveats

Standards-based grading and student performance assessment are promising practices. Still, the path forward is filled with obstacles and hazards. One obvious challenge is practical. Developing common assessments across different education contexts won't be easy. Can expectations be the same at all schools, despite large differences in demography and culture? And what are the potential consequences of establishing benchmarks for performance that are either dependent on or independent of context?

A second challenge is technical. Teacher assessment practices and the assessments themselves must be both valid and reliable for such a system to have any currency. Consequently, performance assessment requires not only building capacity among teachers but also constructing assessments and organizing rating teams. In addition, it requires a larger

policy context that doesn't incentivize gaming. Making portfolios the basis for high-stakes decisions, for instance, would undermine much of the potential of performance assessment.

**What if we could rely on teachers' assessments for the information currently provided by standardized test scores?**

A third challenge is political. Given current distaste for standardized testing, defenders of teacher autonomy and professionalism may interpret a push for consistency in assessment practice as a bait-and-switch. Simply the questioning of teachers' grading practices may raise concerns of academic freedom. Many, no doubt, will see even explicitly teacher-driven efforts as top-down mandates, thus requiring a thoroughly inclusive and transparent process.

In short, there are significant challenges to pursuing such work. Yet evidence from the school, district, and state levels indicates that teachers can evaluate student work accurately and consistently. And although such work may not be easy, our collective will to do what is difficult should be bolstered by the unambiguous inadequacy of the status quo.

## A way forward

There's no one best way forward in seeking to overhaul how we assess student academic achievement. That said, we offer four guidelines for schools, districts, states, and consortia interested in pursuing this work:

- ***Collaborate.*** This work in many ways represents uncharted territory, making for some fairly steep learning curves. Collaborative partnerships, however, can facilitate knowledge sharing. Those pursuing standards-based grading, for instance, can engage teachers, schools, districts, and even states in developing common assessments, calibrating academic expectations, and sharing more accurate grading practices. Technology can bring people together across space. Still, much of this work must be done face-to-face.

- ***Start somewhere.*** Given the enormous challenge of this work, it's important to start with something that feels achievable. School faculty might start by reading and discussing articles on performance assessments and standards-based grading, engaging small groups of teachers to design common assessments and try new grading practices. Slowly, they might scale up, deciding as a school what "proficient" means at each grade level and how to most accurately report a student's performance. At the collaborative, consortium, or state level, cross-district teams might begin work by designing and field testing a small set of standards-aligned, curriculum-embedded performance tasks, and then calibrating student work in cross-district scoring sessions. Whatever the details, the important thing is to get the ball rolling.

- ***Go slow.*** As we know from work at San Leandro High and in New Hampshire, teachers can become invested champions of accurate and reliable assessment practices when given support and guidance. However, these examples also suggest that the process is time-consuming; if it's not supported adequately by the district or state, it will place undue burden on teachers and schools. States interested in pursuing such work should secure adequate resources and begin with small pilot projects. At the local level, a district could convene a task force of teacher representatives and identify pilot classrooms or schools to prototype more calibrated, complex assessments of student performance.

- ***Don't stop halfway.*** Classroom-based data are superior to standardized test data if teachers validly measure student performance and develop ways of reliably reporting on it. New Hampshire started with measurement, while San Leandro High School began with reporting. Despite different starting points, however, the common goal remains to go all the way, as it must be for any group trying to fulfill the potential of teacher-led assessment.

## Getting the data that matter

The newly minted Every Student Succeeds Act (ESSA) extends the mandate of standardized testing and the use of data to make high-stakes decisions. As a result, public pushback against testing is likely to grow in both breadth and intensity. Yet ESSA also invites states to apply for waivers that would let them use locally created assessments for the purpose of determining student competency and school progress, opening the possibility of a third way.

Current standardized testing practices are highly problematic. But testing will be particularly difficult to displace if critics continue to stand only in opposition, offering no replacement. What we're suggesting, then, is an alternative system, one that would rely on teacher assessments that display some of the statistical properties valued by supporters of standardized tests. Such a system might provide the same information captured by standardized tests — about student proficiency and comparative school performance — yet in much greater depth while also getting a great deal more.

With the adoption of teacher-generated performance assessments, we also hope to see the unintended consequences of educational measurement minimized. Narrowing of the curriculum has reduced learning in arts, history, science, health, and other aspects of a diverse curriculum, and teaching to the test has rendered school less engaging for both students and teachers. In addition, hours devoted to testing and test preparation can eat up weeks of instructional time, an opportunity cost that's particularly detrimental to students who need academic support. These consequences, although unintended, need not be accepted.

Finally, places like San Leandro High School and New Hampshire have suggested a number of additional benefits related to this work. Teachers can increase their professional agency and connect more with one another around shared practices. School and district administrators, with more reliable and valid internal evidence, can better target their resources to the students who need them most. Students can focus less on deciphering each teacher's unique expectations, experience lower levels of anxiety, and maintain a stronger sense of ownership over their own growth. Policy leaders can learn more about student learning and school performance. And parents can get better information about what their children are actually learning in school. In short, the upsides are tremendous.

Teacher-led assessment, of course, won't solve the misuse of data. If policy leaders are intent on stigmatizing schools or punishing them for their demography, a different source of student achievement data won't stop them from doing so. Ultimately, though, we must begin by shifting the conversation away from a false choice — between data and no data — to one about how to make data useful. Much of the information we already have is quite valuable. It just needs to exist in a form that's meaningful to those beyond a single classroom.

That's the future of student achievement data. Not one test to rule them all. Rather, many classroom-based assessments and grades that mean something to everyone. K

**We're suggesting a system that would rely on teacher assessments that display some of the statistical properties valued by supporters of standardized tests.**

***Join the conversation***

*facebook.com/pdkintl*
*@pdkintl*

Thinkstock/iStock