thinkpoint

How might we measure what matters most?
It's time for an assessment overhaul

by Jay McTighe

The abrupt and unprecedented disruptions to education brought about by the Covid-19 pandemic have resulted in considerable changes to business as usual in K-12 schools and universities. We have witnessed an almost overnight shift to online learning, home schooling, virtual professional and curriculum development, increased use of open source resources, web-based testing for Advanced Placement courses, suspension of final exams, and alternative approaches to traditional grading. Just as experts have forecast that societies will be fundamentally changed as a result of the pandemic, it also seems likely that schooling as we knew it will never be the same.

A fundamental question must therefore be raised concerning the alignment between our high-stakes assessments and the goals of a modern education: Are we currently assessing everything that matters, or only those things that are easiest to test and least expensive to score?

Indeed, the aftermath of times of crisis inevitably offers an opportunity to step back and reexamine all facets of life, and it is thus an opportune time to scrutinize our current educational system. This paper will focus on one of the most impactful elements of today's education—our current approach to large-scale accountability testing in the U.S. It highlights several noteworthy deficiencies of the present system and will then propose a more comprehensive assessment system that can address these weaknesses and measure more of the learning outcomes that matter most in a modern education.

The present assessment system

While standardized tests have been used for decades, the No Child Left Behind (NCLB) federal law enacted in 2001 raised the bar for the use of standardized testing for K-12 public school accountability. NCLB required all states to conduct annual testing of students in grades 3-8, and one grade in high school, in reading and mathematics. Educational accountability was accomplished by publishing the test results, comparing schools and districts, and enacting consequences for schools that failed to achieve "annual yearly progress" quotas. While school improvement grants were provided to low-performing schools, continued poor performance on accountability tests resulted in school "takeovers," mandated private tutoring of students,

Many individuals within, and outside of, the educational establishment recognize that the present assessment system is flawed, and point out that, ironically, the current accountability mechanism may actually impede the very efforts needed to realize many important educational goals of a modern education.

(Darling-Hammond, 2013; Jimenez & Sargrad, 2017.)

and/or allowing parents to choose alternative schools.

In 2015, the Every Student Succeeds Act (ESSA) replaced NCLB as the nation's major federal education statute. This law also requires states to measure and report on public school performance. However, ESSA's accountability system is less punitive than NCLB's, allowing local schools and districts to take the lead in school improvement, rather than have consequences applied by the state. ESSA also calls for schools to focus on preparing students for "college and career" readiness, and to date, a majority of states have included a broader array of educational outcomes beyond mastery of core academic content in traditional subject areas. These include skills in *critical thinking* and *problem-solving, collaboration, communication, civic and community engagement*, and *social emotional learning*. (See Mishkind, 2014).

While the overall requirements have evolved, the primary measure of educational accountability remains rooted in the results of annual standardized tests. Many individuals within, and outside of, the educational es-

tablishment recognize that the present assessment system is flawed, and point out that, ironically, the current accountability mechanism may actually impede the very efforts needed to realize many important educational goals of a modern education. (Darling-Hammond, 2013; Jimenez & Sargrad, 2017.)

A constricted system

The deficiencies in the present testing system have been well documented over the years. Part of the critique centers on the format; i.e., the nearly exclusive use of a selected-response (primarily multiple-choice) format for test items. Given the large-scale administration of these tests, it is no wonder that they employ this format to enable inexpensive, machine scoring and relatively quick return of results. While multiple-choice tests provide broad, standardized measures yielding comparable results (at least within states), they are not well suited to assess a number of key educational outcomes. For example, virtually all current standards in English language arts (ELA) include listening and speaking skills, which are generally acknowledged as the foundations of literacy. Yet those skills are rarely, if ever, assessed on large-scale accountability tests.

To put it more starkly, important academic learning outcomes are falling through the cracks of the current standardized assessment system. Selected-response assessments (or even brief-constructed responses) are simply incapable of measuring students' abilities to address open-ended problems and issues, engage in discussion and debate, write for genuine audiences and purposes, conduct sound research and experimental inquiry, or devel-

To put it more starkly, important academic learning outcomes are falling through the cracks of the current standardized assessment system. Selected-response assessments (or even brief-constructed responses) are simply incapable of measuring students' abilities to address open-ended problems and issues, engage in discussion and debate, write for genuine audiences and purposes, conduct sound research and experimental inquiry, or develop and critique arguments, yet these are surely vital outcomes.

op and critique arguments, yet these are surely vital outcomes. Furthermore, the so-called 21st century skills of critical and creative thinking, collaborative teamwork, multi-media communication, and use of information technologies are typically not tested on today's accountability measures. Accordingly, they are less likely to receive instructional emphasis.

A fundamental question must therefore be raised concerning the alignment between our high-stakes assessments and the goals of a modern education: Are we currently assessing everything that matters, or only those things that are easiest to test and least expensive to score? Unequivocally, our current standardized testing system fails to assess many of the most valued goals of a modern education.

Consequential validity

High stakes assessments have consequences. In other words, their effects on curriculum, instruction, classroom assessments, and student motivation matter. Indeed, the adage, "what gets measured signals what is important," rings true in education. Students regularly ask their teachers, "will this be on the test?" If the answer is "no," they are less likely to pay attention to it. Large-scale assessments hold similar sway. Teachers and administrators pay close attention to what is tested on state assessments since their results can have high stakes consequences. If something is not assessed, it can quickly diminish in importance and receive less

instructional emphasis. The adage applies to the current crop of accountability assessments required by ESSA.

Given the reality that repeated poor school performance on state measures can result in loss of accreditation, staff and administrative transfers, and lower property values in a community, it is no wonder that educators (especially in low-achieving schools) are incentivized to focus on what is tested and

disregard those standards (and even entire subjects) that are not. The result is often a *de facto* narrowing of the curriculum. Furthermore, the pressure to improve performance on once-a year accountability assessments has prompted well-intentioned teachers and administrators to fixate on the format of the tests and institute a variety of misguided "test prep" interventions. Not surprisingly, we have witnessed an entire cottage industry of off-the-shelf test prep materials that implicitly promise that using them will boost students' test performances.

While the temptation to adopt a test prep curriculum is understandable given the stakes, such an action reveals a fundamental misunderstanding—the belief that the best way of improving accountability test scores is to practice the multiple-choice test format (McTighe, 2017). An overreliance on materials that mimic the format of state tests mistakes the measures for the goals. Such test prep is the educational equivalent of practicing for your physical exam

in order to improve your health! Sadly, the use of classroom time in many schools (at least in the tested grades and subjects) would lead one to conclude that the mission of schools is to improve test taking savvy and raise test scores rather than to strive for meaningful learning on outcomes that matter.

Note: Of course, it makes sense to familiarize students with test format, since selected response format can be an effective assessment method for certain outcomes and students will encounter this format through-

out their school lives. However, an over reliance on "multiple-choice" teaching and practice testing are not the best long-term strategies for developing a well-rounded, educated person or even improving scores on annual accountability tests.

Student motivation and engagement should not be overlooked when considering the impact of high stakes tests. Most learners are not stimulated by superficial content "coverage" (just in case it may be tested), rote learning, skill drills, and test prep worksheets; and when students are bored by their schoolwork, the consequences are well known — they exhibit a minimal-compliance attitude, they act up, or they drop out (figuratively and literally). A related casualty of the widespread use of multiple-choice practice tests and associated teacher-made assessments has to do with a worrisome lesson that this format suggests about learning; i.e., that the goal of school is to figure out the "correct" answer from a set of provided options. Is that a life-lesson that we really want to impart?

Given the acknowledged limitations of large-scale, accountability testing, what changes in our assessment system will make it more likely that we are assessing all outcomes that matter? How might an assessment system promote more authentic and meaningful learning, not just provide comparable measures?

The need for multiple sources of assessment evidence reflects a fundamental psychometric requirement—to allow valid inferences to be drawn, an assessment must align with, and provide an appropriate measure of a targeted goal.

From snapshot to photo album

To begin the exploration of an enhanced assessment system, consider an analogy: testing as photography. The current accountability system takes the form of annual standardized tests in reading and mathematics, and in some states, writing. The results of these once-a-year "snapshots" provide a few pictures—their scores are informative and can reveal patterns of achievement on certain learning outcomes. However, no single photo can provide a complete portrayal. What is needed is the equivalence of a photo album containing a variety of pictures taken over time. Just as a photo album provides more information than any one or two pictures within, the same is true for assessment.

Essentially, assessment is an inferential process. The validity of any assessment has to do with the extent to which its results enable sound inferences about what students know, understand, and can do. Since all forms of assessment are susceptible to measurement error, our inferences are more dependable when we consider multiple sources of evidence. Thus, to be able to draw sound inferences, especially for high-stakes accountability purposes, we need a photo album containing a range of photographic evidence, not just a few snapshots of certain outcomes.

The need for multiple sources of assessment evidence reflects a fundamental psychometric requirement: to allow valid inferences to be drawn, an assessment must align with, and provide an appropriate measure of, a targeted goal. Given that there are different types of learning goals—factual knowledge, basic skills, conceptual understandings, complex processes, and dispositions—we need an associated variety of assessment types to gather valid evidence on a variety of outcomes. To continue the analogy,

our assessment photo album will include pictures taken with a wide-angle lens; e.g. 25-60 multiple-choice items that sample widely from a given domain of knowledge and basic skills. However, our album should also include "close up" photographs that probe a particular area more deeply; e.g., development of an argument or a research process.

How might a qualitative change to the current assessment system address its recognized shortcomings and the negative effects of current high stakes measures? The assessment framework proposed here offers an educationally viable approach for achieving three interrelated goals:

1) assessing the most important educational goals in appropriate ways;

2) providing the specific and timely feedback needed to improve learning; and

3) supporting curriculum planning, and local assessment and instruction for meaningful learning.

A three-legged stool

In brief, what is recommended is a "multiple measures" approach to educational accountability based on a framework consisting of three inter-related components for assessing core standards and other important educational outcomes such as 21st century skills: a) content-specific tests; b) a series of content-specific and transdisciplinary performance tasks; and 3) a local assessment component.

This framework can be implemented nationally, through a consortium of states sharing the same items and tasks (i.e., components # 1 and 2), or on a state-by-state basis. In the event that states persist in using single, annual tests, this multi-measure assessment system can be modified for use at the district level. Each of the three assessment components is described below, and Appendix A summarizes this proposed assessment system in chart form.

COMPONENT #1: Content-specific tests

The first component will be familiar to educators and the general public. It features content-specific tests consisting of selected-response (SR) and brief constructed-response (BCR) items designed to measure particular aspects of content standards. Most current state tests and the National Assessment of Educational Progress (NAEP) use SR and BCR items from which inferences about learning are drawn. These types of tests have proven effective and efficient at sampling a broad array of basic knowledge and skills drawn from academic standards.

It is recommended that these tests be computer-based in order to take advantage of enhanced item types made possible through technology-enabled assessments (for example, see Tucker, 2009), and to provide nearly immediate feedback in the form of detailed item analyses (not just scores). It is further proposed that a Matrix Sampling approach be employed as a

cost- and time-saving means of obtaining accountability information at the school and district levels without subjecting every student to testing every year on every aspect of the Standards. However, states or school districts could opt for census testing if individual student scores are desired.

COMPONENT #2: Content-specific and interdisciplinary performance tasks

Selected-response and brief constructed-response item formats are limited in what they can appropriately assess. To properly assess conceptual understanding, transfer, and more complex skills, we need greater use of authentic, performance-based measures in which students are asked to: 1) apply their learning to a new situation; and 2) explain their thinking, show their reasoning, and justify their conclusions. Authentic tasks call for students to apply their learning in genuine, "real-world" contexts. Accordingly, they are better suited to assess more complex aspects of core standards, such as mathematical reasoning, scientific investigation, and argumentation, as well as transdisciplinary 21st century skills issues involving design thinking and technology applications. Authentic tasks are like the game in athletics. While the players have to possess knowledge (the rules) and specific skills (dribbling), playing the game also involves conceptual understanding (game strategies) and transfer (using skills and strategies to advantage in particular game situations).

Assessing what matters must include assessing performance in a "game" in addition to tests of requisite knowledge and skills, which can be efficiently assessed through the first component described above.

The nation has a history of implementing performance assessments on a large scale. State assessments in multiple subject areas were conducted in Maryland, Connecticut, New York, California, Vermont, and Kentucky, and through the New Standards Project during the 1990s (e.g., see Guskey, 2020 and Ferrara, 2009.) Moreover, we have numerous district, state, and national models of judgment-based scoring of student performance, including state and district-level writing assessments, Advanced Placement® (AP) tests, music adjudications, and International Baccalaureate® (IB) portfolio reviews in the visual arts. Other nations (e.g., Great Britain) include assessments scored by teachers as a major element of their national assessments. These examples demonstrate the efficacy of performance evaluation when the following conditions are established and effectively enacted:

- clear scoring criteria embedded in rubrics,
- sufficient training of scorers,
- anchor examples linked to the performance levels in rubrics, and
- inter-rater reliability protocols.

The performance assessments will be set in real-world contexts and include both content-specific and transdisciplinary performances. It is recommend-

ed that a national database of performance tasks and companion scoring rubrics be established from which national, regional, or state assessments would be generated. In fact, many of these tasks and rubrics can be obtained from existing sets, such as the performance tasks curated by the Stanford Center for Assessment, Learning and Equity (SCALE) and other websites. Additional ones would be developed and certified by teams of experts.

Given the fact that performance tasks require more time to administer, a Matrix Sampling approach is strongly recommended. For example, in a large-scale

of the curriculum at designated time periods during the school year. This provision underscores the importance of linking assessment evidence to what teachers are expected to teach and will make it more likely that students will be prepared to tackle them.

This component of the proposed assessment system reflects the idea that performance-based assessments are, in fact, necessary to fully honor the standards by engaging students in application, as required by the Practices in Mathematics and Science, the Anchor Standards in ELA, and the Inquiry Skills of Social Studies. These dimensions of

---

What is recommended is a "multiple measures" approach to educational accountability based on a framework consisting of three inter-related components for assessing core standards and other important educational outcomes.

---

writing assessment, all students in a grade level would be expected to write, but each student would address only one of three writing genres in a given administration—narrative, expository, or argumentation. There will be sufficient number of students assessed in most cases to enable warranted inferences about general writing achievement at the school level and, potentially, at the classroom level.

A significant feature of this second component is that the performance tasks will be implemented by teachers as part

the standards call for students to "do" the discipline—to perform with their learning—and performance assessments are the proper way to reliably assess them.

A significant challenge to large-scale performance assessment lies in the costs of scoring. It is important to note that in this proposal, the scoring will *not* be contracted to commercial test companies, although companies may be enlisted to help with training, moderation, and reporting. Indeed, a central feature of this plan calls for scoring of

the performance tasks to occur at regional scoring sites and be conducted by teams of teachers on designated professional days. State education departments and regional services agencies will be responsible for the organization, training, and monitoring of the scoring process to ensure that consistent and reliable evaluation occurs. As a practical matter, schools and districts will be expected to align their academic calendars to the scoring schedule to ensure teacher participation during the allocated professional days.

Teachers who have participated in scoring student performances, be it through state/district writing assessments, for AP or IB programs, in conjunction with project-based learning, or via professional learning communities (McTighe, 2008), regularly comment on the value of the experience. Indeed, the high-impact professional learning that accrues when teachers work in teams to score student work needs to be factored into the equation. In other words, the costs of scoring the performance assessment tasks are conceived, justified, and budgeted as a joint expenditure for assessment and professional learning.

An important side benefit of involving teacher teams in scoring occurs as teachers share ideas and resources for addressing students' misconceptions and performance weaknesses revealed during their scoring experience. Emerging ideas for needed instructional interventions will be collected and

compiled in an Internet database, accessible to all teachers in the nation, region, or state. (A similar data base currently

Performance-based assessments are, in fact, necessary to fully honor the standards by engaging students in application, as required by the Practices in Mathematics and Science, the Anchor Standards in ELA, and the Inquiry Skills of Social Studies. These dimensions of the standards call for students to "do" the discipline–to perform with their learning–and performance assessments are the proper way to reliably assess them.

exists for science education—http://assessment.aaas.org/topics.) In sum, the process of teacher-based scoring not only influences the overall costs of performance assessments, it has the potential to positively impact classroom instruction for the good of learning (Goldberg and Roswell, 1998; Goldberg, G., 1993).

In considering the consequences of large-scale accountability assessments, let us not forget the students. In addition to their psychometric purpose, performance assessments can be motivational to students. Since the tasks will be set in more authentic contexts than typical test items, they are more likely to be seen by students as relevant and worthwhile. Moreover, because performance assessments are open-ended and do not generally have a single, "correct" answer, they offer opportunities to allow appropriate "voice and choice" for students. In other words, it is

possible to gather evidence of worthy outcomes without requiring every student to do the same exact thing in exactly the same way. Maintaining a system of high standards does not require absolute standardization of all measures.

**COMPONENT #3:
Local assessments**

A standardized national or state assessment system is incapable of fully assessing each student on every important standard and related educational goal (e.g., 21st century outcomes or the arts) for logistical and cost reasons. Even if it were feasible and affordable, it is unwise to limit accountability assessments to only those measures imposed from the outside. There is a need to include local assessments to allow appropriate measures of locally valued educational outcomes in all subject areas and to permit greater personalization than possible through external, standardized tests and tasks.

Performance standards are ultimately achieved at the local

level. A comprehensive and effective national/state accountability system needs to include a district/school-level assessment component, and initiate policies and incentives to ensure that this local assessment becomes more credible, rigorous, and self-correcting. An analogy from athletics explains how this principle already works in the world

of swimming and track and field. State officials do not have to officiate at every local meet to be assured that the times and distances recorded by the local coaches are sufficiently accurate. There need only be local meets open to the public where the rules are followed and the scoring is transparent, backed by a system of regional and state meets, recorded by official scorers.

This third component of the assessment system is built upon the same logic; i.e., legitimize the role of local assessment by trusting educators with the responsibility of scoring student work in all subject

areas. Make the results, framed in terms of standards, public. Then, verify local scoring through a variety of regional and state auditing systems.

The local component of the assessment system allows for a wide variety of possibilities, including common course exams, independent studies and exhibitions, student passion projects, and interdisciplinary projects involving student collaboration. More specifically, it:

- can appropriately assess important achievement targets (e.g., oral reading and speaking, applications of technology, collaborative teamwork, performances in the visual and performing arts) that may otherwise "fall through the cracks" of the first two components;
- is based on local curricula so that teachers, students, and parents will be more likely to "own" the measures and the results;
- offers greater flexibility and potential for differentiation

(e.g., allowing students appropriate "voice and choice" of topics or products) than will the standardized assessments in the other components;
- honors the tradition of local control of education by allowing local decision making, rather than having all prominent assessments imposed from the outside; and
- supports student accountability, i.e., the results become part of local grading and reporting. (Thus, local report cards should have a section in which grades are provided on performance related to content standards along with profiles of performance on 21st century skills development.)

A cornerstone of this third component is a Student Performance Portfolio—a systematic collection of assessment evidence related to core standards and other important educational goals. The performance portfolio would:

- contain results from the **performance tasks** (described in Component #2);
- contain the results of **the content specific tests***(described in Component #1);
- contain results from the **local assessments**;
- **allow students to contribute evidence** of worthy accomplishments, including evidence obtained outside of school;
- include **longitudinal (i.e., developmental) rubrics** in each subject area to guide judgments about

student achievement *and* enable more systematic **tracking of growth** (i.e., progress toward meeting standards); *

- ■ **be audited on an annual basis by regional-wide teams of educators** and citizen-experts, with two content areas sampled each year; and
- ■ **be examined on a sampling basis by the state** in an audit of the quality of local and regional assessment.

* Note: The external test data will never be reported alone, but as a part of the overall Portfolio profile

Unlike a typical rubric used to evaluate student performance on a specific task or assignment, it is recommended that student performance be judged against longitudinal rubrics based on developmental continua in various subject areas. For examples,

What's needed now is the political will and a systematic plan to achieve this vision: to measure what truly matters and leverage assessments in ways that can promote meaningful learning.

see the American Council of the Teaching of Foreign Language Proficiency Guidelines (2012) and the National Writing Project Analytic Writing Continuum (2010). Such a system has been in place for in Great Britain for all subject areas. Longitudinal rubrics enable educators, parents, and students to track progress over time toward meeting exit standards.

The Student Performance Portfolio and its associated rubrics will serve as the repository of a "body of evidence" of achievement and growth over time.

Like a photo album, it provides a more complete and accurate portrayal of a learner than does any single test score ("snapshot"). It enables "triangulation" of data from multiple sources, ultimately yielding more credible (rich, varied, thorough) assessment evidence of core standards and 21st century skills. Once in place, the portfolio will enable students to graduate from high school with a resume of accomplishment compiled over their school career, rather than simply a transcript of courses taken, "seat time" logged, and a cumulative GPA.

Conclusion

This proposed 3-part system will provide a more comprehensive assessment of learning outcomes that matter, while averting many of the acknowledged problems of current accountability testing. Unquestionably, this ambitious vision will require a "selling" phase to introduce the idea to educators, policy makes, parents, and the general public. Even if the vision is understood and embraced, there will be the need for considerable coordination between state education departments, regional agencies, and local districts and schools to organize the implementation system, enact necessary training for teacher-based scoring, and develop a system for synchronizing and reporting the results.

Predictably, there will be many objections as to why we cannot or should not change the existing system, especially from test companies with a huge financial stake in maintaining the status quo. Psychometricians will also weigh in, citing the difficulties of ensuring scoring reliability of open-ended tasks or the generalizability of their results. Teachers' unions may object to requiring teachers to participate in regional scoring sessions. And policymakers may simply want quick and inexpensive measures to use in rating schools and be unwilling to tolerate the cost or time needed to ramp up to the proposed assessment system. Unscrupulous administrators and teachers will likely figure out new ways to "game" the system, especially if the accountability stakes remain high and threaten their positions or salaries.

When faced with any fundamental change or disruptive technology, the tendency of systems is to lean toward inertia. Nonetheless, when it comes to large-scale testing systems, the old adage is apt: I*f you keep doing what you've always done, you'll keep getting what you've always gotten.*

In sum, educators know the learning outcomes that matter most in a modern education. They recognize the inadequacies of the current accountability testing system in providing proper measures of all valued outcomes. Together, let's envision an improved system, such as the 3-part plan outlined in this paper. What's needed now is the political will and a systematic plan to achieve this vision: to measure what truly matters and leverage assessments in ways that can promote meaningful learning.

Reflection Points

1. This ThinkPoint proposes three inter-related assessment components as a multiple-measures approach to educational accountability. How could such an approach boost access to a quality K-16 education for special populations such as Black, Indigenous, and other people of color (BIPOC); students with disabilities; English learners; and students living in poverty?

2. How can you envision implementing and using the paper's three inter-related assessment components in a traditional face-to-face environment? In a remote learning environment? In a hybrid learning environment?

3. What pragmatic alterations in practice could local education agencies (districts, public school academies, ISDs/ESAs) adopt now—alterations that would not require permission or waivers?

References

American Council for the Teaching of Foreign Languages. 2012. ACTFL Proficiency Guidelines. Hastings-on-Hudson, NY: ACTFL Materials Center.
Retrieved from: https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Darling-Hammond, L. (2013). Developing Students' Higher-order Thinking. ASCD Express, 8 (18).
Ferrara, S. (2009). The Maryland school performance assessment program (MSPAP) 1991- 2002: Political considerations. Paper presented to the National Research Council, December 10, 2009.

Goldberg, G. (1993). Scoring MSPAP: A teacher's guide. Baltimore: Maryland State Department of Education.

Goldberg, G. & Roswell, B. (1998). Perception and Practice: The Impact of Teachers' Scoring Experience on Performance-Based Instruction and Classroom Assessment. Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

Guskey, T. (2020). The Dark Side of Assessment Literacy: Avoiding the Perils of Accountability. AASA Journal of Scholarship & Practice. Vol. 17, No. 1, Spring 2020.

Jimenez, L. & Sargard, S. A New Vision for School Accountability. Center for American Progress. April 2017. Retrieved from: https://www.americanprogress.org/issues/education-k-12/reports/2017/03/03/427156/a-new- vision-for-school-accountability/

McTighe, J. (2008). Making the Most of Professional Learning Communities. The Learning Principal, 3(8), 1, 4–7.

McTighe, J. (2013). Learning to the core: Assessing what matters most. Midvale, UT: School Improvement Network.

McTighe, J. (April 2017). Beware the Test Prep Trap. [Blog Post] NEWSELA.
Retrieved from: https://blog.newsela.com/blog/2017/7/25/u2tlzoasa0gur7ua7uij7ihtop6dkp

Mishkind, A. (2014). Overview: State Definitions of College and Career Readiness. Washington, DC: American Institutes for Research College & Career Readiness & Success Center, 2014). Retrieved from: http://www.ccrscenter.org/sites/default/files/CCRS%20Defintions%20Brief_REV_1.pdf

National Writing Project. (2010). NWP Analytic Writing Continuum. Berkeley, CA: National Writing Project.

National Association of Colleges and Employers. (2016). Job Outlook 2016.
Retrieved from: http://www.naceweb.org/s11182015/employers-look-for-in-new-hires.aspx

Tucker, B. "Beyond the Bubble: Technology and the Future of Student Assessment." Education Sector Reports. February 2009.
Retrieved from: https://www.michiganassessmentconsortium.org/wp-content/uploads/beyond-the-bubble.pdf

To learn more

The Michigan Assessment Consortium website contains a rich library filled with sharable assessment resources as well as online learning modules designed to increase assessment literacy among all education stakeholders. Explore these resources, then spread the word!

Visit www.michiganassessmentconsortium.org.
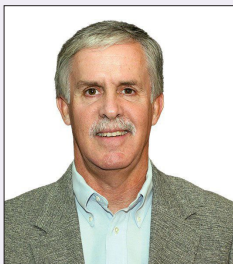
Author

Jay McTighe is an accomplished author, having written more than 50 articles/blogs/chapters and co-authored 17 books, including the award-winning and best-selling *Understanding by Design* series with Grant Wiggins. Jay has an extensive background in professional development and is a regular speaker at national, state, and district conferences and workshops.

Appendix A — A Summary of Features, Costs and Benefits

Assessment Component	Potential Benefits	Potential Drawbacks	Costs
1. Content-specific Standardized Tests ■ selected-response and brief constructed-response formats ■ generally de-contextualized items	■ able to sample a broad array of knowledge and skills within Core Standard areas ■ quick and inexpensive scoring and reporting ■ familiar test format ■ items can be drawn from existing banks (e.g., state tests, NAEP, NWEA) ■ allows for computerized testing ■ standardization allows for comparable results ■ can be used for school/district accountability	■ can encourage the overuse of ■ de-contextualized "test prep" at the expense of meaningful learning ■ may lead to a narrowing of the curriculum (i.e., focus only on the tested content) ■ cannot fully measure important learning areas (e.g., mathematical reasoning, critical thinking, extended writing, research) ■ tests are generally not known in advance	■ comparable to current standardized testing programs* * A national testing program (à la NAEP) would be more cost-effective than mounting 50 different state programs. * A matrix-sampling model could be used to reduce costs (but at the expense of providing individual student scores on every test).
2. Content-specific and transdisciplinary Performance Tasks ■ open-ended ■ require extended constructed responses ■ allow for contextualized and authentic application ■ tasks are scored at regional scoring sites by practicing teachers ■ require rubrics, anchors and inter-rater protocols for reliable scoring	■ able to provide more valid measures of important learning (e.g., mathematical reasoning, critical thinking, extended writing) in greater depth ■ able to assess learners' understanding through contextualized (i.e., more genuine) application, including interdisciplinary contexts ■ 21st Century Outcomes (e.g., technology use, collaborative skills) can be integrated with academic knowledge ■ tasks can be drawn from existing banks (e.g., SCALE) ■ "practicing" for the tasks can support meaningful learning ■ more transparent (i.e., basic tasks and scoring rubrics are known) ■ standardized rubrics and scoring procedures allow for comparable results	■ less able to measure a breadth of knowledge and skills ■ time-consuming to give and score ■ expensive to score ■ judgment-based scoring may compromise reliability ■ delayed results due to time required for scoring	■ Cost estimates can be obtained from several states (MD, CT, KY) that have implemented large-scale performance assessment programs, as well as from many more that conduct state-wide writing assessments. * The costs of scoring the performance tasks should be viewed as expenditures for both measurement and professional development of teachers.

Appendix A — A Summary of Features, Costs and Benefits

Assessment Component	Potential Benefits	Potential Drawbacks	Costs
3. Common Local Assessments ■ allow for a variety of assessment types (e.g., course exams, Senior exhibitions, portfolio collections) ■ based on local curricula ■ can be used for student accountability and local grading ■ features a Student Standards Folder to serve as a repository of achievement evidence ■ scored against developmental continua (longitudinal rubrics) ■ not standardized outside of a school or district, so cannot be used for state, district, or national comparisons	■ allow for a variety of assessment types (e.g., course exams, Senior projects, portfolio collections) aligned to local curricula ■ promote local options and greater "ownership" of measures and results ■ allow for assessing important learning goals that otherwise "fall through the cracks" of the standardized assessments (# 1 and 2) ■ provide more immediate and credible feedback ■ encourage curriculum fidelity and focused instruction ■ can allow for differentiation and student "voice and choice" (e.g., on products) ■ yield individual student scores; can be used for student accountability (e.g., grading) ■ track progress along developmental continua toward meeting standards	■ results are not comparable beyond the school or district ■ not suitable for use in school/district accountability	■ Costs would be dependent on the nature of the curriculum and the chosen assessment options. In general, these costs would be assumed by the local school/district budget.

Jay McTighe answers your questions

1. *Our School Board has adopted a Profile of a Graduate that identifies a set of trans-disciplinary competencies – Critical Thinking, Creativity, Collaboration, Communication, and Self-directed Learning. We value these competencies, but none of them are properly assessed by the standardized tests our state currently uses. Can your proposed assessment system help us gather data on these valued competencies?*

JMcT: Yes! One of the underlying ideas in this paper is the recognition that important educational outcomes, such as those identified in your Graduate Profile, are "falling through the cracks" of the present testing system. The argument for an expanded approach using multiple measures is meant to enable the assessment of more valued educational outcomes, while concurrently supporting authentic learning, and allowing greater personalization for students.

If there were statewide agreement on such competencies, then the 2nd component of the proposed assessment system – agreed-upon performance tasks – could be used to gather some evidence on these. However, I think that the 3rd component – local assessments – would be the most appropriate arena for assembling evidence over time (perhaps in a digital portfolio) of the kinds of competencies identified in a district's Graduate Profile. While the results of local assessments could not be used for state accountability purposes (unless their states established waiver provisions for which their assessments would qualify) a well-developed system of local performance-based assessments can yield:

- performance data that will complement the achievement information obtained through more traditional standardized tests;

- collected evidence on Graduate Profile competencies for digital portfolios;
- greater "voice and choice" opportunities for students; and
- support for meaningful, authentic learning.

In sum, if you identify any competencies as important to the mission of schooling, then you should have a plan to collect assessment evidence about how well your system is working to help students develop them.

2. *While we like the ideas in the paper, especially the increased use of performance-based assessments, we are a small district and quite frankly do not have the resources to develop a comprehensive performance assessment system. How might we realistically do this?*

JMcT: The recommendation to develop and implement a system of high-quality performance tasks and associated rubrics is admittedly a tall order, even for large districts with considerable resources. In states that have regional service agencies (e.g., ESCs, BOCES, IUs,) or for districts having university partners, these organizational entities can play a significant role in supporting the ideas in the paper. In states or regions without an existing entity, my suggestion is to look for opportunities to form consortia of districts within a region or state in order to "divide and conquer" the challenge. Indeed, Assessment Consortia that have been established over the years in Michigan, Maryland, New York, Missouri, and Alberta (CN) provide successful models that can be replicated. Such agencies or consortia can coordinate:

1. the collection and vetting of quality performance tasks and rubrics to share with their member districts;
2. professional development related to implementing and scoring performance assessments;

3. regional scoring sessions for teachers and administrators using established protocols for inter-relator reliability and calibration;
4. the selection and annotation of examples of student work ("anchors") tied to the levels of scoring rubrics;
5. the interpretation of scores on the performance assessments;
6. sharing of instructional ideas and resources to address areas of identified weaknesses in student performance; and
7. sharing of local assessments (the third component of the multi-dimensional assessment system) with member districts as well as with a network of other regional service agencies.

By working smarter and collaboratively, even small districts can realize the benefits described in the paper.

3. *The coronavirus pandemic of 2020 resulted in the suspension of state accountability tests and many educators were relieved. Why are you proposing more standardized tests?*

JMcT: Public schools consume a considerable amount of public funds and taxpayers and policy makers have a right to know how the schools are performing. At present, educational accountability is determined primarily by the results of once-a-year standardized tests relying on a selected-response item format. This paper is based on the assumption that public schools will continue to be held accountable using some standardized measures, and reflects an IF—THEN logic: If external assessments continue to be used as the primary measure of school accountability, then let's make sure that those measures are assessing the right things in the right ways. An expanded approach using multiple measures offers the potential for assessing more valued educational outcomes, supporting more authentic learning, and enabling greater personalization for students.

4. *As the diversity of our student population increases, we are being asked to "personalize" education for all of our students. However, the first two elements of your proposed assessment system are standardized. Isn't that incompatible with the goal of greater personalization?*

JMcT: While some degree of standardization is necessary to allow for comparability of results across schools and districts, an overhauled assessment system as described in this paper offers the opportunity to allow more ways for more students to demonstrate their learning. In particular, the local assessment component will:
a) allow greater "voice and choice" options within performance tasks and student-directed projects;
b) permit educators to collect more personalized forms of evidence; and
c) enable local assessments to establish contexts that align more closely with community experiences and cultures than do external standardized tests.
d) allow for more personalized tracking (and celebration) of individual student progress, gauged against developmental continua rather than only grade-level standards.

Moreover, such an expanded assessment system will likely influence instructional practices by encouraging more authentic learning and differentiated instruction. We can, indeed, maintain high standards without excessive standardization.

5. *What resources can I explore and share with others to begin to deepen understanding on this topic?*

JMcT: The short article, Three Key Questions on Measuring Learning, examines three essential questions: 1) What really matters in a contemporary education? 2) How should we assess those things that matter?

3) How might our assessments enhance learning that matters, not just measure it?

6. *How can you envision implementing and using the paper's three inter-related assessment components in a traditional face-to-face environment? In a remote learning environment? In a hybrid learning environment?*

JMcT: This paper is based on the assumption that public schools will continue to be held accountable using some standardized measures. Accordingly, it argues for an expanded approach using multiple measures as a means of assessing more valued educational outcomes, supporting authentic learning, and enabling greater personalization for students.

While some version of performance-based assessments can be employed in remote and hybrid-learning environments, it is unlikely that these would be sufficiently standardized to allow comparisons for accountability.

Resources for further study

- **Leading Modern Learning: A Blueprint for Vision Driven Schools**
  2nd ed., by Jay McTighe and Greg Curtis. ASCD/Solution Tree 2019.
  https://bit.ly/3fKvPZi

- **Designing Authentic Performance Tasks and Projects: Tools for Meaningful Learning and Assessment**, by Jay McTighe, Kristina J. Doubet, and Eric M. Carbaugh. ASCD 2020.
  https://tinyurl.com/seq83he

- **Demonstrating Student Mastery with Digital Badges and Portfolios**,
  by David Niguidula. ASCD 2019.
  https://bit.ly/3hgjcFR

- **"Three Key Questions on Measuring Learning,"**
  by Jay McTighe. Educational Leadership, February 2018. Vol. 75, No. 5.
  http://bit.ly/2V0b4RR

Relevant MAC resources:

- **How do we design assessment systems for modern learning?**
  https://bit.ly/LP-modern-learning

- **What is collaborative scoring? Why can it be so valuable?**
  https://bit.ly/LP-collaborative-scoring

- **Addressing assessment issues: Value of using a two-component accountability assessment system**
  https://bit.ly/TP-addressing-issues