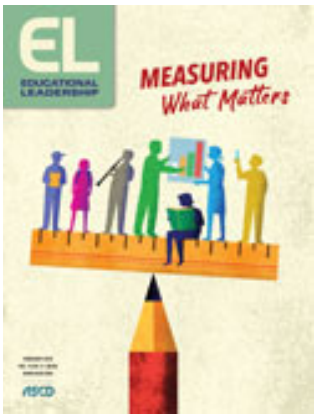
EDUCATIONAL LEADERSHIP

Home | Current Issue | Archives | Buy | Contact

February 2018 | Volume **75** | Number **5**
Measuring What Matters Pages 70-75

Issue Table of Contents | Read Article Abstract



Mission Possible: Measuring Critical Thinking and Problem Solving

Doug Wren and Amy Cashwell

To gauge complex skills, a Virginia district has worked to hone a series of performance assessments.

In 2009—the same year articles in an *Educational Leadership* issue on "Teaching for the 21st Century" recommended that schools assess key 21st century skills—our school district in southeastern Virginia began creating a large-scale performance assessment to gauge students' critical-thinking and problem-solving skills. The following year, nearly 10,000 students in Virginia Beach City Public Schools took the Integrated Performance Task we developed, and hundreds of teachers throughout the district began scoring students' open-ended responses. This was the beginning of a long-running "performance" of our district's performance assessment system—one that continues to this day.

Why did our school system boldly go where few districts had gone before? Because our strategic plan focused on "teaching and assessing those skills our students need to thrive as 21st century learners, workers, and citizens" (Virginia Beach City Public Schools, 2008). And, although we'd created instructional opportunities for students to acquire 21st century skills, we had no way to measure students' performance on these skills districtwide.

We discovered that, although educators have taught critical thinking and problem solving for centuries, assessing these skills *en masse* was more difficult than we anticipated. But we also discovered that developing instruments to measure such skills is possible—and can inform instruction in ways that enhance our ability to teach these skills.

Setting the Stage

While we were developing our new strategic plan in Virginia Beach in 2008, Harvard scholar Tony Wagner told us about an innovative performance assessment for high school students called the College and Work Readiness Assessment (Council for Aid to Education, 2007). After field-testing this assessment, we adopted it as an annual measure of our high school students' critical-thinking, problem-solving, and writing skills. We also decided to create similar performance tasks to administer to all Virginia Beach students in grades 4 and 7. These became our Integrated Performance Task (IPT). Our district was determined to move away from multiple-choice testing and the "deliver and recall" teaching methods it tends to foster.

Act I: Developing Rubrics

To define what critical-thinking, problem-solving, and written communication skills would look like, we developed a rubric spelling out what these skills should involve at the 4th and 7th grade levels. Our rubric employed a 4-point scale (*novice*, *emerging*, *proficient*, and *advanced*), with *3* defined as meeting the standard and *4* as exceeding the standard (Arter & McTighe, 2001).

We reviewed literature and other rubrics aligned with each skill, and then sat down to operationally define critical thinking (CT), problem solving (PS), and written communication (WC). We learned from our mistakes in this process. On an early draft, each skill was subdivided into two or three components —for example, critical thinking was made up of CT1, CT2, and CT3. We soon realized that with this arrangement, test responses would have to be scored seven times! The simpler one-page rubric we ended up with included only CT, PS, and WC.

Figure 1 shows the general operational definition we identified for each skill. As we created specific performance tasks for the Integrated Performance Task, we further defined what the performance of each skill at different levels of this rubric would look like for each task. For instance, Figure 2 spells out what students should be able to do at different levels of critical thinking for one of the 4th grade performance tasks, which involved evaluating an advertisement.

FIGURE 1. Definition of Key Skills on the IPT General Rubric

Skill	Operational Definition
Critical Thinking (CT)	Decides if the information is correct and believable.
Problem Solving (PS)	Makes a choice and gives reasons for the choice.
Written Communication (WC)	Presents information and ideas that are clear, organized, detailed, and written for the intended audience.

FIGURE 2. Partial Rubric for Critical Thinking Skill

(Critical thinking is operationally defined as "decides if information in the IPT booklet is correct & believable")

Level 4 Advanced	Level 3 Proficient	Level 2 Emerging	Level 1 Novice
Gives many examples of information that is incorrect, misleading, or unbelievable. Gives clear, logical, and detailed reasons why each example is incorrect, misleading, or unbelievable.	Gives two or more examples of information that is incorrect, misleading, or unbelievable. Gives clear and logical, reasons why each example is incorrect, misleading, or unbelievable.	Gives one or more examples of incorrect, misleading, or unbelievable information. Gives reasons why each example is incorrect, misleading, or unbelievable. Reasons may be unclear.	Gives no examples of incorrect, misleading, or unbelievable information OR Gives one or more examples but does not explain why information may be incorrect, misleading, or unbelievable.

Note: These descriptions summarize the basics of what a scorer would look for in a test taker's responses to score for critical thinking. The complete rubric provides detail on what to look for and sample responses at each level. For more information on the rubrics associated with our IPT, visit www.vbschools.com/schools/testing/IptHowTo.asp. Used with permission of Virginia Beach City Public Schools.

Act II: Creating Engaging Tasks

One reason we chose the College and Work Readiness Assessment as the basis for our performance tasks for elementary and middle learners is that it's an engaging test. Students have said they like the scenarios involving challenging, real-life problems that this assessment includes (Wagner, 2008). For each task, the assessment provides students with documents—like news stories, editorials, research briefs, and email threads—that give them context for each scenario. Students have authentic-feeling information to consider before they develop a solution to the problem within the task. We emulated these features in our Integrated Performance Task.

We generated age-appropriate scenarios to use as performance tasks, using the GRASPS framework developed by Grant Wiggins and Jay McTighe (2005). GRASPS stands for goal, role, audience, situation, product, and standards for success. Figure 3 shows how we defined each element on the GRASPS framework for a grade 4 performance task.

FIGURE 3. Elements of a Grade 4 Performance Task, Following "GRASPS" Guideline

Element	Description
Goal	To recommend a project to improve students' health at Smith Elementary School and support the recommendation with reasons from the documents
Role	A fourth-grade student at Smith Elementary School
Audience	Mr. Beach, the principal of Smith Elementary School
Situation	A local business will donate money to the school to pay for only one project—either an outdoor fitness course or a fruit and salad bar
Product	A persuasive letter to Mr. Beach recommending one of the projects
Standards	Described in the rubric

Students executing this performance task first see this passage outlining the situation:

> You are a 4th grade student at Smith Elementary School. A local business wants to give your school money to help improve health for all of the students. The money will be used to pay for only one of these projects: An outdoor fitness course at the school or a fruit and salad bar for the lunchroom. Some students want a fitness course and some want a fruit and salad bar. … The school cannot have both. Your principal, Mr. Beach, wants you to help him make a choice.

With this task, test takers receive a fact sheet outlining children's playground injuries, a news story on the benefits of fruit and salad bars, and an advertisement exalting outdoor fitness courses. They receive these prompts:

1. Look at the advertisement on page 5. Find all the information that is incorrect, unbelievable, or misleading. Explain why you think the information is incorrect, unbelievable, or misleading. … Give reasons why the information is incorrect, unbelievable, or misleading.

2. Write a persuasive letter to Mr. Beach explaining your choice for improving health for all students at Smith Elementary School … Use information from this booklet to help you write your letter. [The prompt goes on to list specific elements to include in the letter.]

Again, we learned as we went. Early drafts of our performance tasks were long and wordy and included five open-ended prompts. Realizing that a test's content validity can be compromised if extraneous

variables such as excessive length and readability are added to the mix (Benson, 1981), we reduced the number of prompts along with the length and reading level of material accompanying each task. We further reduced the possibility that reading ability would affect the results by instructing 4th grade teachers to read the directions, scenario, documents, and prompts aloud at the start of the testing period while students followed along in their booklets.

The Virginia Beach district has administered the IPT to our 4th and 7th graders twice a year for more than seven years. Every different performance task has undergone numerous revisions based on reviews and feedback from students, teachers, and assessment experts (including Marc Chun, formerly of the Council for Aid to Education, creators of the CWRA). For instance, when we piloted the Improving Health performance task, a few students said they didn't pick the salad bar because they hated salad. We changed "salad bar" to "fruit and salad bar" and added a photo to show the many food choices a fruit and salad bar offers.

Act III: Finding the Right Scoring System

Two concerns that many districts may have about performance assessments are the potential cost (Picus et al., 2010) and fears that the scoring process might be so subjective that the results will be neither reliable nor valid (Lane & Iwatani, 2016). Our district shared these concerns. As we developed our scoring process, we paid attention to these questions: What will the scoring system cost us—in terms of time and money? How can we make it cost less? How can we make the scores we assign student responses as accurate as possible—including ensuring that different scorers give the same student response a similar score (*interrater reliability*) and that any scorer would give the same student response the same score on a different day (*intrarater reliability*)?

Until recently, we used one method to score student responses on the IPT administered in fall and another to score the spring assessment. Teachers at each individual school scored the fall responses (after some minimal training) and spring responses were scored centrally by a more thoroughly trained cadre of teachers. Employing different methods was appropriate because each assessment served a different purpose: The fall IPT is meant to introduce students to a low-stakes performance task and give teachers formative data they can use to shape instruction. The spring assessment is used more summatively; students and parents see their individual scores, and the district uses the aggregate results to measure its progress on strategic goals.

As with developing the performance tasks, we improved our scoring methods as we went. We realized quickly that, because there wasn't much time during the fall to conduct training sessions at schools, inconsistent scoring between teachers was inevitable. We believed using a centralized scoring plan for the spring IPT would increase interrater reliability on that assessment. But our first effort at centralized scoring showed we had a lot to learn.

One good decision we made was to have each response scored independently by two teachers, with a third teacher breaking the tie if the scores didn't match. Other parts of our initial attempt failed miserably. Our first scoring cadre met in summer 2011 to score the IPT assessment given that spring, and nearly 200 teacher-scorers came and went for four weeks. Bringing in a different group every week and training each scorer to evaluate all three IPT skills was a mistake.

The next summer, we conducted training on the first day of a three-week session. Teachers were required to come on that first day and attend for at least two weeks. Although these requirements reduced the number of scorers, they improved interrater agreement. It also helped that we began training each scorer to focus on only one skill for a single grade level. Scorers never had to shift their mindsets from critical thinking to problem solving to writing skills while scoring a response.

Data and personnel management were also problematic during our first centralized scoring adventure. The following year, we promoted key individuals to manage the training and the data, and assigned one teacher as a supervisor to guide a group of teachers in each of six scoring rooms. Training became more consistent, data was entered accurately, and teacher scorers preferred being supervised by responsible peers. Except for the first summer scoring cadre, interrater agreement between our teacher scorers has ranged from 66 to 82 percent across the three skills at different grade levels.

Recently, after six years of using trained teachers to score responses, we began using computerized scoring for the fall and spring IPT through a vendor. The Turnitin Scoring Engine uses multiple algorithms to replicate the scoring patterns of our most experienced teacher scorers after the engine has been "trained" by having 500 scored student responses fed through the system. This process now makes possible computerized scoring for each performance task scenario. When we develop new scenarios (as we did with one grade 7 task in fall 2017), our teacher scorers start from scratch to "retrain" the system, resulting in new scoring algorithms.

Computerized scoring has demonstrated reliability comparable to what we achieved using human scorers (and above the minimum acceptable value for low-stakes tests) and has cut costs for our spring scoring sessions. Releasing teachers from fall IPT scoring obligations has given them more time to look at their students' responses on the assessment and use what they learn to modify their instruction. However, we realize that scores from any one test seldom tell the whole story. As teachers review their students' IPT results and responses, we suggest they take the advice of Guskey and Jung (2016) and "trust your mind instead of your machine" (p. 54).

Encore: One More "C"

In 2016, Virginia enacted legislation calling for diploma standards aligned with the *Profile of a Virginia Graduate*.[1] The legislation directed the state board of education to give "due consideration to critical thinking, creative thinking, collaboration, communication, and citizenship in the profile" (Virginia General Assembly, 2016). Our IPT was already measuring critical thinking and communication, and we began planning to assess citizenship skills as well. To provide an indicator of these skills, we created new scenarios involving ethical dilemmas that elementary and middle school students commonly face (such as bullying and cheating). We're administering these new performance tasks over the next eight months.

Driving Better Instruction

When we introduced our stakeholders to the idea of the IPT in 2010, it was interesting to see how

different individuals and groups perceived it. Some students, parents, and educators saw it as just another test, but others recognized the value of this new type of assessment. After the initial rollout of the IPT, its value became clearer as we noticed that this performance assessment helped our teachers improve teaching and learning—bearing out what education researchers have found for decades. Many teachers started—or put stronger emphasis on—teaching students to process information, solve real-life problems, and express their thoughts in writing. For example, during the past seven years, social studies teachers have made the shift toward teaching analysis and interpretation of information in document-based performance tasks instead of teaching facts in isolation.

As educators at Virginia Beach schools try to live out the district's mission to prepare all students for college and careers, they now embed authentic tasks and performance-based assessments within every area of the curriculum. While classroom teachers use these smaller assessments to gauge students' acquisition of content as well as 21st century skills, the IPT offers a common, district-level view of our progress at teaching skills deemed essential by our strategic plan. Our teachers continue to use the IPT to gain a better understanding of how their students think and write. There are probably other good ways to assess hard-to-measure skills like problem solving on a large scale. But we think the IPT is a hard act to follow.

EL Online

For a discussion of how to successfully roll out a new assessment system, see the online article "Three Moves for Assessment-System Success" by Jennifer Borgioli.

References

Atkinson, D. (2017). Virginia rethinks high school in its profile of a graduate. *State Education Standard*, *17*(2), 28–33. Retrieved from www.nasbe.org/wp-content/uploads/Virginia-Rethinks-High-School-in-Its-Profile-of-a-Graduate_May-2017-Standard.pdf

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin.

Benson, J. (1981). A redefinition of content validity. *Educational and Psychological Measurement*, *41*(3), 793–802.

Council for Aid to Education. (2007). *College and Work Readiness Assessment* [Measurement instrument].

Guskey, T. R., & Jung, L. A. (2016). Grading: Why you should trust your judgment. *Educational Leadership*, *73*(7), 50–54.

Lane, S., & Iwatani, E. (2016). Design of performance assessments in education. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.), *Handbook of Test Development* (2nd ed., pp. 274–293). New York: Routledge.

Picus, L. O., Adamson, F., Montague, W., & Owens, M. (2010). *A new conceptual framework for analyzing the costs of performance assessment*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. Retrieved from https://scale.stanford.edu/system/files/new-conceptual-framework-analyzing-costs-performance-assessment.pdf

Virginia Beach City Public Schools. (2008). *Compass to 2015: A Strategic Plan for Student Success*. Retrieved from www.vbschools.com/compass/2015

Virginia General Assembly. (2016). Code of Virginia § 22.1-253.13:4. Retrieved from http://law.lis.virginia.gov/vacode/22.1-253.13:4

Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need—and what we can do about it*. New York: Basic Books.

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Alexandria, VA: ASCD.

Endnote

[1] This student profile is the foundation of the Virginia Board of Education's redesign efforts to better prepare our students to participate in the global economy (Atkinson, 2017).

Doug Wren is educational measurement and assessment specialist, and Amy Cashwell is chief academic officer for Virginia Beach City Public Schools.

KEYWORDS

Click on keywords to see similar products:
21st century learning, communication, critical thinking, assessment and grading, formative assessment, standards, whole child: challenged, audience: administrators, audience: district-

based-administrators, audience: higher-education, audience: new-principals, audience: new-teachers, audience: principals, audience: teacher-leaders, audience: teachers, audience: building-level-specialist, audience: instructional-coaches, audience: superintendents, audience: students, level: k-12