Chapter 2

Designing a Standards-Aligned Performance Assessment System

People learn more when they do something.

—Dedranae Tucker, Envision student

Consider the driver's test. Your fellow citizens won't share the road with you until you pass it, proving you can drive skillfully and responsibly. There may be variations to the licensure process, but they all culminate in your getting behind the wheel of a car, an assessor buckled into the seat next to you, and showing you can drive *by driving*.

Compare that to the driver's permit exam, which in most states is a standardized written or electronic test that assesses one's knowledge of the rules of the road. It is a significant step in the process; indeed, passing the permit exam is what allows you to get on the road and practice driving—under the supervision of a licensed driver.

No one questions the distinct purposes of these two assessments. One tests what you know about driving; the other tests how you drive. They are both important, but not equally so. Common sense tells us that a basic knowledge of traffic laws does not provide sufficient evidence to answer the ultimate question at hand: Is the license seeker qualified to drive a car on his or her own? That's why the process culminates with a performance assessment.

Our K–12 education system has much to learn from such common sense. We give our students lots of permit exams and hardly any driver's tests. We measure (or attempt to measure) what they know, hoping it serves as a proxy for what they can do. The upshot is that many young adults leave high school unprepared to drive the metaphoric roads of college or career.

This is not a news flash to most of our readers. We'd rather take up the question that follows: Knowing that performance assessment is a better way to measure and prepare our students, how do we elevate it to its proper role in education, shifting the emphasis of assessment from knowing to doing?

Answers are out there, but they typically direct attention to teachers (here is how their courses should change) or to the larger school system (here is how standardized testing should change).

But the most effective agent for transforming assessment is overlooked. It is neither the teacher nor the state. It is the school.

Performance Assessment Defined---and Refined

Performance assessment is a fancy term for a simple concept: **evaluating what you can do by observing you doing it**.

The musical director auditions singers by hearing them sing. The coach tries out players by watching them play, both through drills (which isolate certain skills) and in scrimmages. The performance doesn't necessarily have to be observed in the doing. A record or product of a performance satisfies in many cases: a photographer's portfolio, a cook's prepared dish.

Performance assessment doesn't have a direct opposite, but it does have counterpoints. A multiple-choice test is the most commonly cited example of what performance assessment is not (Darling-Hammond & Adamson, 2010). For an assessment to merit the qualifier *performance*, the test taker must construct an answer for himself, rather than selecting an answer from predetermined options. By definition, a performance assessment evaluates **a product or performance, requiring some kind of constructive or creative act**.

Another counterpoint is assessment by proxy. When it's not possible to measure someone's abilities directly, we try to measure them indirectly. An example is the traditional job interview. Employers generally can't observe on-the-job performance before they put you on the job. Instead, they sit you down on the other side of the desk and ask you interview questions.

Because your palms are sweating, you certainly feel as though your performance is being assessed. What is really happening is that your ability to do the job is being inferred from proximal data: the confidence and intelligence of your answers, the professionalism of your dress, the details of your reported experience. Most job interviews cannot directly assess your ability to do the job; rather, they measure how well you can interview.

(Exceptions abound, of course, including the “demo lesson” often demanded of applicants for a teaching position.)

Before something can qualify as performance assessment, its challenge must be aligned to its purpose. It is not enough to see someone writing or experimenting or singing or interviewing—in other words, engaged in some creative or constructive activity—and then slap the word *performance* onto its assessment. First, you must name what you want to measure. Only then can you judge whether the observed product or performance gives you the evidence you seek.

This is why defining performance assessment as everything non–multiple choice is simplistic. We can, for example, fool ourselves into thinking that if students are merely writing something—whether a short answer or a developed essay—then we have definitively moved them into the land of performance assessment.

But writing can be as formulaic as any bubble test, especially when it doesn't challenge students to tap a higher-order thinking skill. There's a good reason why English teachers, for example, challenge each other and their students to move beyond the book report as a response to literature. Summarizing plot is not a higher-order thinking skill. If your goal is to assess students' ability to analyze or evaluate text, then the traditional book report, even though it is an open-ended writing assignment, does not qualify as a performance assessment. By definition, **a performance assessment must enact the skill you are intending to measure**.

type="concept"

> To qualify as a **performance assessment**, what is evaluated must be
>
> 1. A product or performance
> 2. An application of a *targeted* skill (or skills)

You know you have a genuine performance assessment if preparing for the test, taking the test, and then applying the skill in real life all look the same. Consider the skill of parallel parking. How is it tested on the driver's test? By parallel parking. How do you practice for that part of the test? By parallel parking. And what do you do with the acquired skill after you pass the test? You parallel park.

An Old Pedagogy for a Newly Demanding World

Long before the term *performance assessment* ever existed, Professor Henry Higgins took Eliza Doolittle to the horse races (Lerner & Loewe, 1956). Proof that Eliza's learning had gone deep was not to be found in the cozy confines of Higgins's library, where she recited “the rain in Spain” until her accent was perfectly aristocratic. The bet was that Higgins could make her pass as an aristocrat; naturally, she needed to prove her skills among them. So off they went to the “Ascot opening day.”

It was a challenging test in a new and authentic context. Of course, in one of the more memorable scenes of musical theater, Eliza, until that moment indistinguishable from the blue bloods surrounding her, exhorts her horse to “move your bloomin' arse!!!” (p. 78). Eliza wasn't quite ready to pass this famous example of a performance assessment.

Jargon deserves our skepticism. So often it turns out to be wrapping paper. You tear it off and are miffed to find an old concept, regifted.

Performance assessment, as a term, can make us feel that way, especially when it is uttered in a tone that pretends to be some revolutionary new invention. But good teachers have, through their own judgment and sense, been designing good performance assessments for as long as humans have been teaching each other things. Many of us can think back to a teacher whose course culminated with some demanding assignment, one that required us to *do* the subject rather than just learn it, one that not only challenged us but *helped us make sense of what the course was ultimately about*. Student-designed experiments, research projects, presentations of learning—such oft-cited examples of performance assessment were around long before the term gained currency. Visual and performing arts teachers have always dwelt in the land of performance assessment, never left, and have good reason to wonder what all the hubbub is about.

Still, performance assessment, though not describing a new thing, is an increasingly useful term, for two reasons. One, a lot of non–performance assessment has grown up around it, competing for sunlight. The denser the thicket, the more attentive we must be to the difference between what to prune and what to let grow.

Two, though performance assessment has always been good pedagogy, it is fast becoming the only pedagogy that can possibly address the demands of this changing world. Tony Wagner, an education professor at Harvard, has been a consistently eloquent voice on the matter:

> Today, because knowledge is available on every Internet-connected device, what you know matters far less than what you can do with what you know. The capacity to

> innovate—the ability to solve problems creatively or bring new possibilities to life—and skills like critical thinking, communication and collaboration are far more important than academic knowledge. As one executive told me, “We can teach new hires the content, and we will have to because it continues to change, but we can't teach them how to think—to ask the right questions—and to take initiative.” (Friedman, 2013)

That job falls to our schools. And that list—thinking critically, communicating, collaborating—keeps popping up every time someone thinks about preparing students for their future, whether for college, career, or citizenship. It is a set of skills, not a body of facts. It's a list of verbs, not nouns. If we agree on their importance, then we must design educational experiences that allow students to practice and teachers to coach those skills *in action*.

A multiple-choice question is not inherently bad. It can be an appropriate and certainly efficient way of assessing certain kinds of content knowledge. But when cast in the light of the task at hand, it just looks woefully inadequate, even irrelevant. We realize that performance assessment—however jargony it may sound—is our only possible means for measuring what our students need measured.

Actually, there is much to appreciate about the term performance assessment, because the words it comprises point in the direction that school design needs to go. The word *performance* connotes action, creativity, and the presence of an audience. The word *assessment* suggests something ongoing—a process—in a way that the word *test* does not. These are important themes for the design of a schoolwide performance assessment system.

The Envision Performance Assessment System

The linchpin of Envision's Deeper Learning Student Assessment System is the portfolio defense, described in chapter 1. We see high school as a four-year project, everything building toward that one final performance.

But as we explained in the previous chapter, that one final performance in fact brings together four performances:

- A research paper
- An analysis
- An inquiry
- A creative expression

This list emerged from a decade of rigorous dialogue, not only among our teachers but also with educational experts at Stanford, who themselves have carefully studied the articulation between high school and college in the United States. This is our organization's distillation of college and career readiness (and it therefore subsumes the Common Core State Standards as well). We have planted four flags in the ground: if our students can master each of these tasks, they can succeed in college.

Each of the performance tasks is guided by its own scoring rubric, used by students and teachers to determine whether a given performance has missed, met, or exceeded the standard of proficiency.

Much of the professional development at Envision Schools centers on how to implement these performance tasks and use their associated rubrics. A weeklong session orients new teachers to the system. August pre-school planning requires all teaching teams to map out the number of opportunities students will have to perform these tasks. And over the course of the year, during both weekly collaborative meetings as well as dedicated days of professional development, teachers use the rubrics to score and discuss student work together, a never-finished process of interpretation and calibration. It then becomes the job of principals and lead teachers to ensure that the tasks are taught frequently enough and deeply enough so that, by the end of four years, students have had a chance to attain mastery of each of these skills.

By the junior and senior years, students are actively working with their subject area teachers to craft their best possible research papers, analyses, inquiries, and representations of creativity. Senior year, they choose the best of each, then revise it to proficiency if it is not already, or polish it to an advanced level if it is already deemed proficient. When ready, the artifact is submitted, reflected on, and finally defended.

Key Features of the System

Over the years, the design of our system has evolved, and as long as we remain committed to being a learning organization, it will continue to evolve. Still, certain features have developed into design principles that are holding fast.

The List of Performance Assessments Is Short

It has to be.

A commitment to performance assessment—whether from a teacher, a school, or a school system—is a commitment to focus. The number of performances considered central must be kept small, ideally no more than the fingers on one hand.

There are two reasons for this. The first is practical. Learning is a process of repetition, and repetition can only occur if there is time for it. Having twenty priorities is the same as having none. Performance assessment is ultimately about goal setting, and we all know that one of the biggest barriers to reaching goals is having too many of them.

The second reason is pedagogical. One of the fundamental purposes of education is to help us make sense of a complex world. We can't reduce that complexity, but we can learn how to navigate it. We do this by discerning patterns, building systems, and molding theories.

A well-designed performance assessment is exactly that: a theory of action that focuses the learner and the teacher, unifying a complex web of skills and content into a comprehensible whole.

The Performance Assessments Distill the Standards

Standards setting is inherently an act of analysis. A set of standards is always an answer to the question, “What is quality?” Naturally, we seek the answer by taking the subject in question and breaking it down into its parts.

What makes for a great tennis player? Few of us would be satisfied with the answer, “A player who wins a lot of games,” even though it's true. The purpose of the question is to parse the performance: the mechanics of the backhand and forehand; the speed of the serve; the player's fitness, grit, and grace under pressure. Whether it's a performance or a machine, we take something apart when we want to see how it works. And after we're done, we end up with a lot of parts spread across the floor.

So it's unfair to complain, as many often do, that academic standards come to us as long, overwhelming to-do lists. Even standards that attempt to condense and prioritize, as the Common Core standards do, fill pages and pages with their discrete items, broken down by skill areas, broken down by grade level, and so on. If it were an engine and you wanted to see how it works, then a quality education would have many parts to spread across the floor.

So we describe quality with analysis; in contrast, we *assess* quality with synthesis—or at least we should. Assessment is most convincing when we can see how all those parts come together. Many athletes look great on paper, analytically, but greatness is ultimately achieved on the court, in the game. You can survey all those engine parts spread across the floor—all of them accounted for and gleaming—but you don't know if you have a quality engine until you put it together and turn it on.

Yet we make this mistake in our schools all the time. We approach standards as though they were a checklist, testing kids on discrete skills and tidbits of knowledge without ever asking them to synthesize and apply. Education becomes a practice session toward a game that is never scheduled. It's an engine tinkered with in the garage that never hits the street.

We shouldn't blame the standards for this. It's not the job of standards to tell us how to put the engine back together. In fact, most standards consciously avoid doing so; often they preface themselves as the *what* instead of the *how*. The how is the job of educators. (And truth be told, we wouldn't have it any other way.)

This is the purpose of a performance assessment system: to take all the various goals vying for attention, from all realms—state standards, district initiatives, college entrance requirements, school mission, academic traditions, and 21st century skills—and synthesize those into a few key performances, whose achievement convincingly makes the claim, “This is quality.”

When Envision undertook the work of distilling all of its educational goals, we ended up with four fundamental skills to master: writing a research report, making an inquiry, conducting an analysis, and actualizing a creative vision. Within the first week of her first year at school, a ninth grader knows that these are the four things that she must learn how to do well. And for the next four years, all of her courses, assignments, lessons, and projects—the component parts of school—feed into her mastery of those skills. In the end, there is a unified assessment, the defense of her portfolio, that challenges the student to *put it all back together*.

None of the Performance Assessments Is Tied to a Particular Subject Discipline

At first glance, the subject disciplines—English, math, history, and science—appear absent from our list of performance assessments. In fact, they are baked in, because these are the courses in which students learn to accomplish these tasks. Particular tasks naturally emerge from particular subject disciplines. But—and here is the advantage of a list that is not discipline specific—it is a one-to-many, rather than a one-to-one relationship. The responsibility for teaching how to perform each of these tasks is, in every case, shared. This design embodies the zeitgeist of the Common Core era, which proclaims that the teaching of literacy must be shared across subject areas.

Moreover, by untethering a research paper from, say, history, we can prepare students for the reality that certain modes of discourse will show up in subject disciplines that are not taught in high school, that are encountered for the first time in college, and that many of our students will go on to major in: sociology, political science, geology, and so on.

Still, it is important to note—especially for the proud defenders of academic tradition—that although our design may sometimes blur the distinctions between traditional subject disciplines, it does not erase them. In fact, we've observed the opposite: the more the subject disciplines collaborate together, the more we are learning of the important differences in subject-specific thinking. Historical research, for example, is intrinsically different from literary analysis, though both require close and careful reading. Such differences are not merely respected; they are nurtured and celebrated, and they often serve as fodder for impressive reflections during portfolio defenses.

The Rubrics Are Commonly Shared

A rubric is not merely an assessment tool; more fundamentally, it is a communication tool. Using words instead of numbers or symbols, a rubric serves to explain what it means to do a good job.

For that communication to be most effective, a rubric should be used both before the performance ("What is expected of me?") as well as after ("How did I do?"). The same rubric should also be used across multiple performances, offering many chances to meet one clearly articulated set of expectations. Mastering a skill comes not only through practice but also through a deepening understanding of the expectations. For this to happen, the learner needs more than one opportunity to demonstrate progress in relation to the same expectations.

The more opportunities, the better. The benefits of a rubric's repeated use within one course only compound when the rubric is used across multiple courses and multiple years. At Envision, we have designed rubrics that work across "grade bands"—ninth/tenth and eleventh/twelfth—allowing students to work with a given rubric for two years.

If a school community is converging on a set of performance assessments, then it should also converge on an associated set of common rubrics. Envision's rubrics are shared across our whole network of schools. You can find samples of Envision's rubrics in the appendix.

Designing Performance Assessments

A complete performance assessment has three parts:

1. **The outcome(s)**

   Designing a performance assessment begins with announcing what you hope that the learner achieves, specifying the targeted skills or standards to be measured. Often, the outcomes are framed as "learning targets" or "objectives."

2. **Demonstration of the outcome(s)**

   This is the "task," the "assignment," or the "prompt"—what the learner is asked to do, resulting in a product or performance that provides direct evidence of the targeted skills or standards.

3. **Measurement of the outcome(s)**

   The criteria for success must be established before the learner creates the product or delivers the performance. Typically, this is documented in the form of a rubric.

type="concept"

> Teachers at Envision are able to focus their design efforts on the task (part 2), because the outcomes (part 1) and the rubrics (part 3) have already been established and are shared schoolwide.

With this three-part structure in mind, let us look at two examples of Envision performance assessments.

A Scientific Inquiry: Disaster in the Gulf

Here is an example of a performance assessment, designed by Envision teachers Stanley Richards and Ben Rosen, that is nested within a larger interdisciplinary project that explores the question, "Who is responsible for the 2010 BP oil spill in the Gulf of Mexico?" Students researched the policy and laws pertaining to the spill in their Government class. In their English class, they researched and wrote first-person accounts of how the spill affected people in the gulf, the oil company, and the government, and delivered these accounts at a mock congressional hearing.

In art class, students created works that interpreted the effects of the spill on nature. For science, the students conducted an experimental inquiry into the best methods for cleaning up the oil spill. Here is an overview of the three parts:

Part 1: The Outcomes

To demonstrate their mastery of the inquiry competency in science, students must complete a performance assessment that embodies the following outcomes:

- Initiating the Inquiry

  What is the evidence that the student can formulate questions that can be explored by scientific investigations as well as articulate a testable hypothesis?

  - Asks empirically testable, scientific questions
  - Constructs drawings, diagrams, or models to represent what's being investigated
  - Explains the limitations and precision of a model as a representation of the system or process
  - Formulates a testable hypothesis that is directly related to the question asked

- Planning and Carrying Out Investigations

  What is the evidence that the student can design and perform investigations to explore a natural phenomenon?

  - Designs controlled experiments (with multiple trials) to test the suggested hypothesis
  - Identifies and explains the independent and dependent variables in the hypothesis
  - Clearly communicates the details of the procedures so that they can be replicated by another group of students
  - Creates a detailed and clear data collection method for all trials
  - Conducts multiple trials

- Representing, Analyzing, and Interpreting Data

  What is the evidence that the student can organize, analyze, and interpret the data?

  - Organizes the data in tables and/or graphs

- Expresses relationships and quantities (units) using mathematical conventions
- Explains mathematical computation results in relationship to the expected outcome
- Analyzes and interprets the data and finds patterns
- Draws inferences from the data
- Suggests strengths or weaknesses in inferences from which further investigation could result

- Constructing Evidence-Based Arguments and Communicating Conclusions

  What is the evidence that the student can articulate evidence-based explanations and effectively communicate conclusions?

  - Constructs a scientific argument, explaining how data and acceptable scientific theory support the claim
  - Identifies a counterclaim (possible weaknesses in scientific argument or in one's own argument)
  - Provides multiple representations to communicate conclusions (words, tables, diagrams, graphs, and/or mathematical expressions)
  - Draws conclusions with specific discussion of limitations
  - Uses language and tone appropriate to the purpose and audience
  - Follows conventions of scientific writing, including accurate use of scientific/technical terms, quantitative data, and visual representations

Part 2: The Task

The Disaster in the Gulf inquiry task challenges students to analyze techniques for removing spilled oil from the water and wetlands. Students research various cleanup solutions, generate a hypothesis, and then create and implement a scientific investigation to determine whether their hypothesis is correct. The task was designed so that students could practice toward and produce evidence of the desired outcomes listed in part 1.

In the appendix, a sample of student work illustrates one student's journey through the performance assessment, from his research to his hypothesis to his experiment using oil, detergent, and cotton balls.

Part 3: The Rubric

In the appendix, you will find the scientific inquiry rubric used to evaluate the student work for this performance assessment. It breaks down the outcomes listed in part 1 and communicates how to measure those skills across four performance levels: emerging, developing, proficient, and advanced. The same rubric is used across multiple assignments over multiple years. Teachers use it to map what they must teach, and students use it to understand what is expected of them and to plot the development of their skills.

type="example"

> Two documents in the appendix provide fuller detail on the Disaster in the Gulf performance assessment:
>
> - Scientific Inquiry Performance Task and Rubric
> - Disaster in the Gulf Student Work and Reflection

A Textual Analysis: Dante's *Inferno*

This performance assessment is also nested within a larger project, called the *Inferno* Mosaic Retelling Project, designed to engage eleventh- and twelfth-grade students in a rigorous reading and analysis of Dante's 14th century epic poem, the *Inferno*. (The mosaic retelling also works with ninth and tenth graders using Homer's *Odyssey*.)

The project revolves around two portfolio-eligible performance assessments: an artistic expression and a textual analysis. (Rubrics for both are included in the appendix.) After the class reads the poem, each student chooses some of Dante's lines to interpret artistically in the medium of her choice. The students present their art publicly as part of an ensemble retelling of the poem. In the final step, each student writes a literary analysis essay based on the lines that she has interpreted artistically.

We'll talk more about project design principles in the next chapter. Here we focus on one of the performance assessments—the textual analysis:

Part 1: The Outcomes

To demonstrate their ability to read and think critically and to communicate effectively, students must complete a performance assessment that embodies the following expectations, which are aligned with the Common Core State Standards for English Language Arts:

- **Argument**

  What is the evidence that the student can develop an argument?

  - Responds to the texts with a controlling idea or argument that demonstrates engaged reading and critical thinking
  - Acknowledges and responds to key questions, concerns, or alternative claims relevant to the controlling idea/claim
  - Makes insightful connections, raises implications, and/or draws meaningful conclusions as a result of the reading and analysis

- **Evidence and Analysis**

  What is the evidence that the student can support the argument and analyze evidence?

  - Examines one or more significant works of fiction and/or nonfiction
  - Examines and analyzes the ideas and points of view presented in the texts and the author's language used to convey those ideas (for example, figurative language, literary elements, rhetorical devices)
  - Provides relevant textual evidence to support ideas and claims

- **Organization**

  What is the evidence that the student can organize, analyze, and interpret the data?

  - Presents the controlling idea/argument in a way that is clear and guides the paper's organization
  - Demonstrates a coherence and an internal structure that supports the argument
  - Consistently uses transitions that relate and connect one idea to another
  - Develops ideas and claims in appropriate depth

- **Conventions**

What is the evidence that the student can use language skillfully to communicate ideas?

- Uses grammar, language, and techniques that are appropriate to the student's purpose and audience
- Observes appropriate language conventions
- Engages the reader with a strong voice and rhetorical technique (for example, anecdotes, “grabber” introductions, repetition, sentence variety, parallelism)
- Cites textual evidence accurately and consistently

Part 2: The Task

In a textual analysis essay, students take the lines from Dante's *Inferno* that they have already interpreted artistically and now interpret them analytically, presenting and supporting a student-generated thesis.

Before developing the paper, each student must propose a thesis to the class for feedback and approval. A structural outline is also required before work can begin on the first draft. These steps help the student address the expected standards of the performance assessment (listed in part 1).

Many students struggle with interpreting another's words—especially a great author's—with words of their own. This is exactly why the art task, perhaps counterintuitively, comes before the writing task. By drawing the students into a different medium of expression, the art task effectively forces the students into an act of interpretation. For the textual analysis paper, students return to the land of words to explain what their art helped them notice.

One student, for example, noticed something surprising after creating a sculpture of the monster-guardian Geryon, made from parts he found in a junkyard. Once it was sculpted, the monster appeared horrifying to the student, yet in his initial reading of the poem's lines, the monster sounded “cool,” not scary at all. Upon further analysis, the student noticed that the narrator describes Geryon, who guards a lower circle of Hell, in a tone very calm and matter-of-fact, unlike the narrator's high-pitched, fearful, connotation-rich descriptions of earlier monsters guarding higher circles of Hell. This observation led the student to a sophisticated thesis linking style to theme: the deeper his journey into Hell, the more calmly, even coldly, Dante the Pilgrim accepts what he sees there, which is exactly what Virgil, and God, expect from him.

Part 3: The Rubric

Student essays are evaluated, both formatively and summatively, using the English Language Arts Textual Analysis rubric (included in the appendix).

type="example"

Video 6. The *Inferno* Mosaic Retelling Project

Watch how Justin's assignment to interpret Dante's *Inferno* through art helped his students gain a deeper understanding of the epic poem and its themes, enabling them to write more perceptive textual analysis papers.[ic02uf001.jpg]

type="tip"

Resources for Designing Performance Assessments

- *Envision Performance Assessment Planning Template (in the appendix)*

Envision developed this tool to help our teachers and our client-partners design performance assessments. The template guides your thinking through all three parts of a complete performance assessment: (1) What are your desired outcomes? (2) How will students demonstrate those outcomes? and (3) How will you measure them?

- *SCALE Performance Assessment Quality Rubric (in the appendix)*

Longtime assessment experts at the Stanford Center for Assessment, Learning, and Equity (SCALE), in consultation with practitioners in the deeper learning community, developed this rubric for evaluating performance assessments. The tool helpfully isolates the various features of a quality performance assessment, including alignment to standards, clarity of task prompt, and level of student engagement. Excellent professional development can be built around this tool; for example, a group of teachers can gather to share designed assessments and use the rubric to give each other constructive feedback.

- *Designing for Deeper Learning: How to Develop Performance Assessments for the Common Core (a free online course: novoed.com/learning-design-common-core)*

Our colleagues at SCALE, who helped Envision design its performance assessment system, introduced a MOOC (massive open online course) on performance assessment design in fall 2014.

> Envision's Campaign Ad Project serves as a featured example in the Stanford course.

The Challenges Are the Strengths

Performance assessment is hard. It is complex and time consuming at every stage, and it requires constant maintenance. Doing it well, beyond the scope of a single classroom, comes with all the challenges of collaboration. As soon as you start to scale it across a department or school or school system, you run into problems of validity and reliability (technical terms from the world of standardized testing).

But the power of performance assessment lies guarded within these very challenges. Only by facing them does the work reach its potential. Here we catalogue the perceived problems and show how each one is a latent strength.

The Challenge: Performance Assessment Is "Costly"

The Upside: The Size of Our Investment Is Equal to the Size of Our Return

Compared to its alternatives, performance assessment is almost always more expensive, time consuming, and resource intensive.

Again, compare the driver's permit exam to the driver's test. One requires a piece of paper or a computer and can be scored by a machine. Draw up the test once, and it can be disseminated to thousands. The other requires one trained human being to sit in a car for up to an hour with every single person seeking a driver's license. In comparison to the permit exam, the driver's test is incredibly costly.

But we do it because we, as a society, have decided that it's worth it. When it comes to putting skilled and safe drivers on the road, performance assessment is what it takes to ensure the outcome that we seek.

Clearly, this book argues that performance assessment in K–12 education is also "worth it." Value is not simply a function of cost; it is equally a product of investment. The more time you invest in something, the more value it holds for you. The more valuable it is to you, the more you care to invest in it. This is a virtuous cycle.

We've seen this time and again in our schools. When teachers see good performance assessment design as one of their primary responsibilities, and when they are given appropriate support to fulfill that responsibility, they treat these performance assessments with great care. A carefully designed, skillfully implemented, and reliably scored performance assessment system requires a significant investment from teachers and school leaders, but that investment creates its own commitment. Students benefit immensely from this.

The Challenge: Performance Assessment Design Is Complex

The Upside: The Result Is Powerfully Simple

Testing companies can churn out multiple-choice questions like widgets from an assembly line.

A good performance assessment, in contrast, can only be produced with hand-craftsmanship. It requires careful thinking through every stage of design, from the targeting of skills, to strategizing how to produce evidence of those skills, to determining how to measure that evidence. Because performance assessments tend to synthesize a range of standards and subskills, the designer must puzzle a number of pieces together. It's hard work.

But when done well, what emerges from that complex wrangling can be powerful in its simplicity. What the learner sees is a coherent, singular goal, a way to apply all the parts of his learning toward a whole: after my months in this course, here is the experiment I can now design, the argument I can now defend, the art I can now create, the math I can now apply to a real-world problem. More powerfully than any other mechanism, a well-designed, unifying, and culminating performance assessment communicates to the learner the *meaning* of what he has learned.

The Challenge: Performance Assessment Tries to Measure Skills That Are Hard to Measure

The Upside: Collaboration and Revision Are Required

The world of standardized testing is nervous about performance assessment because of a concept known as *validity*. A test is valid if it accurately measures what it intends to measure.

Because performance assessments try to measure skills that tend to be complex and hard to quantify (for example, research, analysis, inquiry, and creativity), psychometricians (big word for the people who study the validity of testing) see such assessments as fraught with the potential for error.

But here the difference between hard and impossible is important. Performance assessment is hard—not impossible—to design for validity. And the difficulty can be overcome by increasing two inputs: time and manpower. The greater the number of people involved in a performance assessment's creation and the more time given to its revision, the closer to perfect it can be.

How convenient that a culture of collaboration and revision happens to be the most effective agent for improving teaching and learning. What we've noticed in our schools is that performance assessment design provides a deeply authentic reason for our teachers to come together. As Arthur Costa and Bena Kallick (1995) have written, “Teams build assessment—and assessment builds teams” (p. 141).

The Challenge: Performance Assessment Is Hard to Score Reliably

The Upside: It's the Best Professional Development Ever Invented

An argumentative essay, a historical research project, an extended science inquiry—these are assignments that can't be scored by a machine or with an answer key. They require human judgment.

When humans are making judgments about complex work, then one has to be concerned about an issue that is known in the field as *reliability*. Is the score on this performance based on predetermined, commonly understood, and static criteria, or is it a reflection of an individual and inconstant judgment? If a piece of student work earns a wide range of scores from a group of judges, then psychometricians won't trust any of those scores as “reliable.”

When performance assessments are administered on a large scale—for example, AP tests—enormous effort is invested in establishing “inter-rater reliability.” Many different people are judging the tests, but those people have been trained to look for similar things and reach similar judgments.

It may seem that reliability lies beyond the concern of a particular teacher giving a particular assignment at a school. As long as a student understands ahead of time what is expected of him, and the evaluation of his work follows through on those expectations, then why should it matter whether another teacher down the hall would score the work a little differently?

In a traditionally organized school, the answer to the last question is that it doesn't matter much. In a course credit system, the teacher owes little to the school beyond a letter grade for each student. What that letter means is largely up to the teacher.

But if a school moves toward a performance assessment system, and the orientation shifts from counting credits to mastering skills, then suddenly there is a real need for teachers to reach a shared understanding of “What is mastery?”

Over the last twenty years, the tool that has gained widespread acceptance for meeting this challenge is the rubric, whose defining characteristic is its insistence on words, rather than abstract symbols, to describe the quality of the work.

Educators are well familiar with the rubric's typical format—a table, levels of quality across one axis, various aspects of the performance across another, with cells containing phrases that describe features of the student work. Truth be told, most rubrics disappoint in some way. We're never quite satisfied with the wording (English teachers in particular). If created in collaboration, then they are created through compromise. And at this point, we've all seen enough of them that certain features border on cliché.

So it's not the form of the tool—that blizzard of bullets and boxes—that gives rubrics their power. Rather, it is the practices—the thinking and the actions—that surround rubrics that have made them a transformative force in education. First of all, creating a rubric requires us to do the all-important but often underemphasized mapping backwards from a goal. We must define proficiency, establish the standard. Most rubric writing starts there, filling in the boxes that give words to the expectation we have for all of our students. One of the biggest barriers to learning is lack of clarity around the learning target. Rubrics force us to describe the target, and the benefits that ripple from this act cannot be overestimated.

Second, even if the rubric has been handed down to you, as is often the case with rubrics that are common across a school, the work of calibration never ends. The student work keeps rolling in, so the conversation of what the rubric means is never finished. Just as every Supreme Court case is an attempt to interpret the words of the Constitution, so too is every scoring of a student paper or performance an interpretation of the meaning of the rubric. Teachers need to keep coming together, looking at student work together, and reaching agreement on varying levels of quality, also known as calibrating. Sometimes this leads to revision of the rubric; more often it leads to a refined understanding of students' abilities.

We hear it time and again, after a session during which teachers gathered around a table, using a rubric to score student work together: “That's the best professional development I've ever had.”

Of course, a rubric makes such collaboration possible but not inevitable. The hard drives of most teachers are scattered with the bones of old rubrics, unused and forgotten. A department cooks one up during some August professional development, rolls it out with some energy that September. But by November, it's a check-in agenda item during the department meeting. By March, most of the department are grading their papers without it.

Performance assessment is a powerful engine, but it doesn't drive itself. It needs tracks to a destination. School leaders must make it a priority and build structures that allow teachers to collaborate regularly on their use of schoolwide rubrics. Sustainability comes with building performance assessment into the design of the school.

The Tailwind of the Common Core

Even though performance assessment has always been the right thing to do, teachers and schools must often fight headwinds to get it done. In the last decade of the high-stakes bubble test, those headwinds have never blown stronger.

Many, including Envision Schools, have had to keep a strong commitment to performance assessment when all that “counts” are the bubble-test scores. You want to believe that it's going to work out, that if students are trained in higher-order thinking, then multiple-choice questions are just an interesting puzzle to solve on the fly. Sometimes it works out this way. ELA is an area where Envision's commitment to performance assessment has paid off in terms of test scores. It's a riskier road in the so-called content disciplines (science and history). And in math, it's a fraught endeavor. Train kids to think and do math as it is practiced in the real world, and they will suffer in the bubble tests, which seem incapable of rewarding anything but rote, algorithmic learning.

Refreshingly, the Common Core puts the wind at our back for the first time in a long time. They aren't perfect, but these standards go a long way toward closing the gap between what we have to do and what we should be doing. It has elevated the importance of literacy, shifted the focus to higher-order thinking skills, and reinforced the idea of learning how to learn.

Many of the state standards to which schools have been held in the past have emphasized what students should know over what they should be able to do. The Common Core flips the center of gravity to the other side, from knowing to doing. In ELA, the standards stress the ability to write arguments based on evidence, conduct research, read across the curriculum, engage in academic discussion, make formal presentations, and use technology effectively. In mathematics, standards stress conceptual understanding, applying mathematical thinking to real-world issues and challenges. At the high school level, there is an emphasis on mathematical modeling.

The “content” is still there. The math standards lay out a learning progression from K to 12, starting with whole numbers and addition and subtraction, and moving into geometry, algebra, probability, and statistics. In ELA, there is grammar and Shakespeare. But in reading the standards, you can't help but be struck by the radical shift of priority: content is the means; skills are the end. That's because the ultimate goal of the standards is very clear: college and career readiness. There is an acknowledgment that college is a journey into new content. Skills, not memorized packets of static knowledge, are what a student will need there.

The Common Core State Standards take pains to avoid dictating methods or even recommending approaches to assessment, as they should. But read between the lines. Notice all the attention to higher-order thinking skills. Performance assessment offers the only possible approach to assessing what the Common Core is asking us to do. Even the large-scale standardized tests, normally able to bubble-test everything, can't get around it. Smarter Balanced and PARCC both needed to develop new performance assessment components for their Common Core tests; there was no other way to align them to the standards.

Whatever the limitations of the Common Core, a set of standards that necessitates performance assessment is better than standards that don't. In a nutshell, that is why the Common Core is an advance for education: without any kind of performance assessment, school is a glorified driver's permit exam.

It remains critical, however, that we keep the Common Core in perspective. We must not mistake the winds of policy for the pedagogical port of call.

The Common Core *is* policy, not pedagogy, and the winds will shift again. At the writing of this book, it is already clear that the journey through the Common Core will be a blustery one. The tests will be controversial. Some states will back out. Components will be revised. The tests will get rewritten. And one day, the Common Core will be replaced by something else. This is why schools must establish a graduate profile for their students that transcends any set of discrete standards, including the Common Core.

Take the long view. Treat the Common Core as an opportunity to speed your journey. Maximize its potential to help transform your school. But beware the siren song: in this policy climate, it's easy to start thinking of the Common Core as the destination. When the winds shift—and they will—you risk losing the true course.

In the meantime, we should enjoy this wind at our back, for however long it lasts. The Common Core does validate two ideas that all students deserve: college and career readiness as a goal, and performance assessment as an essential strategy. This is a huge opportunity to make our schools both more rigorous and more engaging for our students.

type="example"

Video 7. The Envision Assessment Process

Envision Education, in partnership with Stanford University, developed performance assessments linked both to standards and to deeper learning skills, so that all Envision teachers use the same rigorous assessment tools. Watch teachers discussing and evaluating student work in collaboration.[ic02uf002.jpg]

References

Costa, A. L., & Kallick, B. (1995). Teams build assessment—and assessment builds teams. In A. L. Costa & B. Kallick (Eds.), *Assessment in the learning organization: Shifting the paradigm* (pp. 141–152). Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education (SCOPE).

Friedman, T. (2013, March 30). Need a job? Invent it. New York Times. Retrieved from http://www.nytimes.com/2013/03/31/opinion/sunday/friedman-need-a-job-invent-it.html

Lerner, A. J., & Loewe, F. (1956). *My fair lady*. New York, NY: New American Library.