Next Generation Assessments for Measuring Complex Learning in Science

Jody Clarke-Midura and Chris Dede, Harvard University
Jill Norton, Rennie Center for Education Research & Policy

Introduction

Assessments can either drive or constrain innovation in education. The development of new assessments that measure student performance against the Common Core State Standards offers a powerful point of leverage in the effort to ensure that the new standards deliver on their promise of enhanced educational opportunities. New assessments can help to focus teachers’ and students’ attention on essential knowledge and skills, and provide more timely and useful information for states, schools, teachers, parents and students themselves. It is vital that these new assessments measure student achievement in more complete, authentic, and meaningful ways.

Currently, the Common Core State Standards exist only for English-language arts and mathematics, but work is underway to develop national K-12 standards for science. The National Research Council (NRC), Achieve, Inc., National Science Teachers Association (NSTA), and American Association for the Advancement of Science are all engaged in efforts to develop new national science standards. While numerous papers have summarized research findings and made recommendations to both assessment consortia on critical assessment issues, none has focused specifically on the possibilities for next generation assessments for assessing science content and inquiry skills.

The limitations of current state assessments are well documented, and are the fundamental reasons for the creation of the two national assessment consortia. Paper-and-pencil item-based tests are intrinsically incapable of providing authentic measurement of the complex intellectual and psychosocial performances that are essential for 21st century work and citizenship. The inadequacies of traditional assessments have been particularly troubling in science. Typical multiple-choice tests are insufficient for determining how well students are developing sophisticated inquiry skills in science—a key capability for science, technology, engineering, and mathematics (STEM) careers. Many reports and studies have documented that higher-order thinking skills related to sophisticated cognition (e.g. inquiry

1 For this paper, we use two definitions of inquiry. The first, is based on White, Collins, and Frederiksen’s (in press) definition, which centers on theorizing, questioning & hypothesizing, investigating, analyzing, and synthesizing. This detailed definition is placed in the context of the National Science Education Standards general definition of scientific inquiry as: “…the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work…also…the activities through which students develop knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world.” National Research Council (1996). National Science Education Standards. Washington, DC: National Academies Press.
processes, formulating scientific explanations, communicating scientific understanding, strategies for resolving novel situations) are difficult to measure with multiple-choice or even constructed-response paper-and-pencil tests (NRC, 2006). Research has shown that these tests also are incapable of showing whether or not science instruction is effective in helping students learn inquiry (Quellmalz et al., 2007). The development of more valid assessments of science inquiry and related skills holds the potential to ensure that students are not only better prepared for the STEM fields that are essential for the nation’s economic viability, but also better prepared for a broad range of work and citizenship responsibilities.

This paper describes the potential of next generation state assessments for science that utilize modern technology for delivery, scoring, and reporting. Its argument also makes the case for new types of assessments aligned to the NRC’s vision for state science frameworks that encourages “students to actively engage in science practices in order to deepen their understanding of core ideas” over time. The paper illustrates these possibilities through a case study of a research project at the Harvard Graduate School of Education focused on developing virtual performance assessments (VPAs) in science. The approach described here aligns with the National Assessment of Education Progress science framework, which emphasizes 1) a paring down of content, 2) assessing students’ conceptual knowledge and 3) performance assessments.

While discussions about using computer adaptive technology and about revising assessments to better assess English learners are ongoing, the development of new assessments for science is in an earlier stage. This paper seeks to help in shaping formative conceptions of science assessments. The hope is to unbind notions of science assessments from what is and to embolden new ideas about what these assessments could be if we truly mean to prepare all students for college and the careers of the 21st century.

Background and Context

Numerous reports on economic development in the context of globalization state that inquiry and complex reasoning are critical skills for competing in our knowledge-based, global economy (NRC, 2006; President’s Council of Advisors on Science and Technology (PCAST), 2010). The worker of the 21st century must have science and mathematics skills, creativity, fluency in information and communication technologies, and the ability to solve complex problems.

Yet, data indicate that the U.S. is failing to promote student mastery of crucial science knowledge, skills and abilities, including inquiry practices and complex reasoning. For example, the 2009 National Assessment of Education Progress (NAEP) results reveal that a majority of the U.S. students who took the test received scores below proficient. Similarly, on the 2007 Trends in International Mathematics and Science Study (TIMSS), only 10 percent of U.S. eighth graders reached the advanced benchmark on the science portion—demonstrating a weak grasp of complex and abstract scientific concepts. On the 2009 Program for International Student Assessment (PISA), only 29 percent of students in the U.S. demonstrated ability to complete higher-order tasks such as those involving scientific explanations. That current instruction in science does not focus on these skills explains some of this lackluster performance; teachers center students’ learning on what is measured in summative tests (Darling-Hammond, 2010). Re-thinking science assessments to better measure inquiry and complex reasoning will help refocus instruction on these crucial skills.

The Nature of Inquiry

Scientific inquiry is hypothesized to be the method by which scientists study the world. In order to promote scientific reasoning, the NRC argues that students must participate in authentic practices of science. As discussed in White, Collins, and Frederikson (in press), while detailed definitions of inquiry can be complex, at its core this suite of processes centers on theorizing and investigating. For example, Kuhn, Black, Keselman, and
Kaplan (2000) define inquiry learning as investigations where students individually or collectively investigate a set of phenomena (virtual or real) and draw conclusions about it.

Below are two of the related skills we measure in the particular virtual performance assessment (VPA) described in this paper:

- Student develops a scientific explanation of what is happening in the virtual world that includes: 1) a claim about the phenomena, 2) the evidence (either empirical or observations), and 3) reasoning that links claims with evidence.
- Student gathers data that help explain or provide evidence to justify the claim being made.

Later, we discuss how our VPA elicits and measures skills such as these in detailed ways more similar to the work of scientists than do multiple-choice, short answer, or essay questions.

Current Challenges in Measuring Inquiry

The NRC recently released a public draft of their conceptual framework for new science education core standards. A primary emphasis in this new framework is that learning about science and engineering involves the integration of knowledge of scientific explanations (i.e. content knowledge) with the practices and skills needed to engage in scientific inquiry and engineering design. Similarly, other national and international science frameworks, such as the Science Framework for the 2011 NAEP, the College Board Standards for College Success, and the Programme for International Student Assessment (PISA) of the Organization for Economic Co-operation and Development all place emphasis on integrated science practices and performances.

The mastery of discrete science facts as measured on most standardized tests is an inadequate representation of whether students know or understand integrated science practices. In response, for over two decades, researchers have sought to develop alternative assessments for measuring science that involve tasks in an authentic or real-life context or that mirror the workplace or other real-life contexts. For example, in the 1990s Maryland used hands-on performance assessments in science as part of their state testing program. Numerous studies were conducted on performance assessments to assess the psychometric properties of these alternate assessments. These studies also focused on the feasibility (i.e. cost effectiveness and practicality) of using these types of measures on a large scale. Research findings indicated that these alternate assessments were more aligned to the knowledge, skills and abilities (KSAs) being measured and more valuable for providing feedback to teachers about ongoing student attainment than multiple-choice tests.

However, there were several limitations to the use of hands-on performance assessments as summative assessments for accountability. Research on hands-on performance assessments found:

- Students performed differently on similar tasks that were supposedly measuring the same construct; ideally, a student would perform consistently on various tasks assessing a construct (Shavelson et al., 1993);
- Students performed differently on identical tasks on different occasions (Cronbach et al., 1997);
- Hands-on performance assessments are cost-prohibitive when compared to multiple-choice tests (Stecher & Klein, 1997); and
- Hands-on performance assessments still have limited validity, despite their authenticity, compared to multiple-choice tests (Linn et al., 1991).

Affirming this last point, a recent study found, “Even the hands-on performance tasks in these large-scale science tests are highly structured and relatively short (15-40 minutes), truncating the investigation strategies that can be measured” (Quellmalz et al., 2007, page 1).

New Opportunities in Technology-based Assessment

Fortunately, since the performance based assessment studies of the 1990s, three advances have taken place that potentially enable online performance assessments capable of validly measuring the full complexity of scientific inquiry: 1) advances in cognitive science, 2) advances in statistics and measurement, and 3) advances in information and communication technologies. To illustrate the power and potential of these new types of performance assessments, this paper describes our
current research on one such model: immersive virtual environments (IVEs). IVEs are three-dimensional (3-D) environments, either single or multi-user, where participants’ digital personae (avatars) engage in virtual activities and experiences. These immersive, interactive media have become commonplace in many people’s lives, through gaming (e.g. World of Warcraft and America’s Army), social interactions and learning (such as Second Life, Club Penguin), and recreation (e.g. The Sims Online, Webkinz). Part of the attraction of IVEs is that they can simulate complex real-world settings in which participants can enact the types of complicated processes that underlie various real-world workplace roles. Both military and medical education have benefited from this capability (Fletcher, 2009; Kneebone, 2005).

Research has established that, when well-designed, IVEs can aid students in learning authentic, sophisticated inquiry practices (Mayo, 2009). Our studies of virtual environments as curriculum provide an example. For almost a decade, our research team has studied the feasibility and practicality of using IVEs to increase student achievement in scientific inquiry (Dede, 2009). In this research, we studied how virtual environments enable students to do authentic inquiry and engage in the processes of science. Our first series of studies, funded by National Science Foundation from 1999-2009, were on River City. The River City curriculum was a multi-user immersive virtual environment (MUVE) designed to teach middle school science (Clarke et al., 2006). The curriculum was centered on skills of hypothesis formation and experimental design, as well as on content related to national standards and assessments in biology and ecology.

We were able to implement MUVE-based curricula in a wide range of schools in a manner that teachers and technology coordinators found practical and scalable. We worked with over 200 teachers and more than 20,000 students. We conducted a series of quasi-experimental design studies to determine if virtual environments can simulate real-world experimentation and provide students with engaging, meaningful learning experiences that increase achievement in scientific inquiry. Using conventional paper-and-pencil, item-based measures, our results from a series of research studies showed that these virtual environments enable students to engage in authentic inquiry tasks (problem finding and experimental design) and also increase students’ engagement and self-efficacy (Clarke & Dede, 2007; Clarke et al., 2006; Ketelhut, 2007; Nelson, 2007). This seminal research on IVEs was discussed both in the U.S. Department of Education’s National Educational Technology Plan (2010) and in the National Research Council’s report on games and simulations in science education (2011).

Even though paper-and-pencil tests captured some of students’ learning, we found that students’ performance on the multiple-choice pre-post-tests typically used as measures in this type of research did not necessarily reflect learning that we saw via interviews, observations, summative essays, and analyses of log file data that capture students’ activity as they interact with the environment (Clarke, 2006; Ketelhut & Dede, 2006; Ketelhut et al., 2007). We built rich case studies of student learning in which we triangulated and compared different data sources, both qualitative and quantitative, in order to illustrate and understand students’ inquiry learning (Clarke, 2006; Clarke & Dede, 2005, 2007; Ketelhut et al., 2007). A finding from our experience was that paper-and-pencil item-based assessments, even after extensive refinement, do not fully capture students’ learning of inquiry skills.

Further, our immersive curricular environments and similar interactive, immersive media enable the collection of very rich datastreams about individual learners that provide better ways to assess inquiry processes (Clarke, 2009; Ketelhut et al., 2007). We believe that these streams of “active” behavioral data on student performances can be utilized in the development of virtual assessments. While research on game-like simulations for fostering student learning is starting to proliferate, studying the potential of this medium for summative assessments of student learning in a standardized fashion is still in its infancy.

The U.S. Department of Education’s 2010 National Educational Technology Plan (NETP) identifies increased global economic competition as a fundamental challenge facing the U.S. over the next decade. The NETP therefore calls for immediate action to ensure that today’s U.S. students are learning 21st century skills that foster innovation and economic prosperity.
One recommendation in this Plan focuses specifically on developing sophisticated forms of technology-based assessment: “2.3 Conduct research and development that explores how embedded assessment technologies, such as simulations, collaboration environments, virtual worlds, games, and cognitive tutors, can be used to engage and motivate learners while assessing complex skills” (U.S. Department of Education (USDOE), 2010, page 19).

In research on assessment, IVEs enable investigators to measure authentic, situated performances that reflect the inquiry processes used by scientists (Donovan & Bransford, 2005).

Our current studies center on whether IVEs can provide reliable, practical, affordable summative assessments in accountability settings that are more valid than paper-and-pencil, item-based tests in measuring science inquiry and similar sophisticated STEM skills. Immersive virtual performance assessments use the IVE interface to offer a new model for how we measure higher-order skills such as problem-solving, causal reasoning, and inquiry learning. Our research is synthesizing over two decades of studies on performance assessments, measurement theory, cognition, technology, and videogame design. In the following section, we provide a case study of our work developing and studying summative virtual performance assessments that measure science inquiry processes as part of a national or state high stakes testing program. This research exemplifies the new types of assessments Partnership for the Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium (SBAC) should explore and promote.

The Virtual Performance Assessment (VPA) Project

With funding from the Institute of Education Sciences (IES), the VPA project at the Harvard Graduate School of Education is developing and studying the feasibility of immersive virtual performance assessments to assess scientific inquiry of middle school students as a standardized component of an accountability program (see http://vpa.gse.harvard.edu). The goal is to provide states with reliable and valid technology-based performance assessments linked to state and National Science Education Standards (NSES) academic standards for science content and inquiry processes, extending capabilities to conduct rigorous studies that provide empirical data on student academic achievement in middle school science.

In order to ensure that we were measuring what we intended to measure (inquiry), we used the Evidence Centered Design (ECD) framework (Mislevey & Haertel, 2006; Mislevy & Rahman, 2009) to design our assessments. ECD formalizes the procedures generally done by expert assessment developers. Using the ECD approach allowed us to articulate every aspect of the assessment from the knowledge, skills, and abilities (KSAs) that they are measuring to the types of evidence that will allow one to make claims about what students know. In addition, we are using the Principled Assessment Designs for Inquiry (PADI) system, which is software for creating assessments for science inquiry based on the ECD framework. Design templates allow assessment developers to create multiple forms of the same assessment. Using these frameworks, we have reframed science inquiry constructs (theorizing, questioning and hypothesizing, investigating, analyzing and synthesizing) into specific KSAs aligned with current national standards. Through the process of articulating the exact details of what is being measured and how it is being measured, it is easy to link the KSAs to evidence of student learning (see Appendix A for a Table describing an extended framework for VPA design and development). Linking KSAs like this provides a measure of validity that research has found often lacking in performance assessments (e.g. Linn, Baker et al. 1991).

Description of the Assessments

Traditional assessments often focus on individual test items and rely on student affirmation as a response that indicates knowledge. In our VPAs, we base the evaluation of student performance on measurements captured as in-world interactions. These interactions allow us to assess what students know and do not know about science inquiry and problem solving. As a part of the inqui-
VPA in Action: An Illustration

To demonstrate how the performance assessment works at the classroom level, the following is a brief description of a VPA.

It is May, and students in Ms. Jones’ eighth grade science class are participating in a virtual performance assessment pilot program. As part of their state accountability program, all students in eighth grade must demonstrate proficiency in integrated science practices. These assessments are summative assessments that are meant to sample the domain of inquiry. Students take the virtual assessment in a block period. The assessment lasts about fifty minutes.

Ms. Jones logs into the VPA teacher’s portal and creates accounts for her students, selecting the initial assessment she wants them to take. When class starts, the students sit at individual computers and login to begin their simulated experience.

Arielle sits at her computer and logs into the student portal. She opens the assessment and is immediately allowed to choose what her avatar looks like. She selects an avatar and enters the world.

Figure 3. A VPA avatar selection screen.
The camera slowly provides an aerial view of the world to orient Arielle to the problem space. Arielle sees that there is a village and what appear to be farms with ponds. The camera then focuses on a multi-colored frog with six legs. Arielle wonders, “What could be causing this frog to have six legs?” The assessment begins. A scientist and farmers who have just discovered this mutated frog greet Arielle. The farmers all offer competing hypotheses for why the frog is mutated. The scientist turns to Arielle’s avatar and tells her that she must conduct an investigation and come up with her own theory, backed up with evidence. He asks her if she thinks any of the hypotheses provided are plausible.

Figure 4. Characters presenting competing hypotheses in a VPA.

Figure 5. Setting up the problem.
Arielle must come up with a claim and support it with evidence and reasoning. In order to make her claim she must first gather data. She sets out to explore the farms.

Just as scientists collect data, Arielle has options of different kinds of data she can collect. There is a lab she can visit to run tests. For example, she can collect water, tadpoles, and frogs from each of the four farms. She can then bring them to the lab to conduct water tests, blood tests, and genetic tests. In addition, prior research studies are available, and residents provide data about their points of view.
Figure 6. A VPA backpack containing a limited number of items.

Arielle must make a choice about what data she thinks is the most important or that she wants to investigate first. One aspect of doing inquiry is knowing what data will justify a claim. We want to examine students’ data gathering strategies and see how they correlate to the claims they make. Thus, we limit the amount of data that students can carry in their backpack. If students were allowed to pick up every piece of data in the world, then it would be difficult to make inferences about their knowledge of what data is important evidence in the investigation. If students were asked to evaluate a piece of data every time they collected it, then the task would become boring.

Thus, the design requires students to make choices through actions. Students can carry only eight pieces of data at a time. They can go to the lab at any time to run tests on the data (e.g. water tests, blood tests, genetic tests). Any piece of data discarded from the backpack will go back into the world and can be picked back up at any time (given there is space in the backpack). This is not a design that is related to game play, nor is it meant to model how many samples to collect. It is a structural feature that applies constraints in order to force students to be more thoughtful about the data they collect in this assessment.

Also, the availability of prior research studies helps to "level the playing field" between students who start with stronger content knowledge and those who begin with weaker. This is important in ensuring that the assessment is measuring inquiry process skills rather than content knowledge.

Arielle collects eight pieces of data from two farms. She realizes that she cannot carry any more and decides to go to the lab to run some tests. She arrives at the lab and examines the water samples. Her tests show that the lab water and water from one of the farms contains pesticides. However, one of the farms has clean water. She runs genetic tests on the two frogs she collects and sees that they are the same. She notes that both of the frogs have high counts of white blood cells. She decides that she needs to learn more about what these tests tell her. She goes to the research kiosk and looks up information on blood tests and pesticides. Arielle started with examining data and then moved to research that would allow her to reason from the data.

At the computer on Arielle’s left, Maria is tackling the assessment differently. Initially, she decides to examine prior research studies. She examines the research that is available on frogs and tadpoles. She reads about viruses and genetic mutations in frogs and decides to gather data in order to determine which is the cause. She goes to each of the four farms and collects a tadpole and a frog to run tests on. Back at the lab, she finds that all of the frogs have similar genetic make-up. However, two of the tadpoles have small tails. She notes that a frog from the same farm also has a virus in its blood. She looks up the virus in the research documents and believes she has found evidence. She speaks to the scientist and builds a claim for why the frog is mutated, including evidence-based reasoning from the research she conducted.

After class, Ms. Jones reviews the reporting tool to see the diagnoses the assessment provides about what each student knows and does not know about gathering data about a claim, making a claim, and then supporting it with evidence and reasoning. The tool presents data at both the individual and class level, and Ms. Jones finds that, while the majority of students are strong in providing evidence, they are weak in reasoning from evidence. Also, some students collect data related to hypotheses they have already discarded, or collect all possible data without seeming to have a hypothesis. This shows weaknesses in their inquiry skills.

This particular VPA is designed as a summative assessment to sample from the domain of science inquiry. Overall, VPAs can serve as one component of a comprehensive science assessment (e.g. an inquiry portion of a summative assessment for science).
ry progression embedded within the VPAs, students are required to make a series of choices as they participate in an ongoing narrative. The focus of the VPAs is not the attainment of a single right answer, but rather on the result of a series of choices that students make. The series of interactions produces rich observations that enable us to make a fine distinction of students’ understanding of the various facets of inquiry discussed earlier.

**Important Aspects of the VPA Model**

The skills we are measuring in this particular VPA focus on gathering data about a claim, making a claim, and supporting it with evidence and reasoning—skills that we argue are difficult to capture in multiple-choice and open-response tests. By setting up the assessment in a virtual environment, we can follow students’ trajectories of data gathering. We then can correlate these to the claims they build and the evidence and reasoning they use to support those assertions. Each challenge in our assessments relies on students collecting data and providing evidence to support a claim, and students’ scores are based on the evidence and reasoning they provide for a given claim.

In our prior research in immersive virtual environments mentioned above, we did not find students’ experience playing video or computer games to be a predictor of their performance in the curriculum. Thus, we hypothesize that videogame and computer game experience will not be a predictor of student performance on our VPAs; we are conducting research in order to test this hypothesis. If we are correct in our assumption, then the only strategies students can learn in order to do better on our VPAs involve applying inquiry practices—the domain processes on which instruction should focus. In contrast, students can learn test-taking skills that aid them in correctly selecting multiple-choice answers even though these skills do not provide additional knowledge of the domain; this distorts what is measured.

We support the concept of multiple modes of assessments that triangulate to form claims about what a student does or does not know. Our early work in virtual performance assessments has centered on measuring students’ ability to reason from evidence, as a demonstration of concept for a much broader set of VPAs measuring a wide variety of knowledge, skills, and abilities important in authentic practices. These can then complement more conventional forms of testing.

Facing a mandate to implement digital assessments by 2014, states are already moving towards computer-based testing because of the types of sophisticated intellectual and psychosocial performances they can measure. Whether these future assessments are merely digitized versions of paper-and-pencil tests or offer the increased features of VPAs, similar infrastructure, resource, and policy issues must be addressed. Fortunately, based on our experiences with scaling up MUVE-based curricula, we believe that the investments made to enable simple digital assessments will be sufficient to enable the use of VPAs. In other words, at scale, the expense involved in constructing and using VPAs is roughly comparable to other forms of high quality testing.

In summary, virtual performance assessments have numerous advantages over hands-on performance assessments:

1. Standardizing the administration of hands-on performance assessment is difficult, so extensive training is required. In contrast, VPAs ensure standardization by delivering instruction to students in an identical manner via the technology.

2. VPAs alleviate the need for developing, shipping, and providing schools with materials and kits for hands-on tasks. All that is needed are a computer and an internet connection.

3. Scoring will all be done by the technology, so no raters are necessary, reducing cost, training, and the possibility of human error.

4. VPAs potentially have fewer problems with task and occasion variability.

VPAs also have advantages over conventional paper-and-pencil tests and digitized versions of these:

1. Multiple-choice, short answer, and essay questions do not present a realistic context within which to elicit the processes of complex performances such as scientific inquiry.

2. VPAs mirror for students the types of inquiry processes to which teachers should orient instruction more accurately than do conventional measures of inquiry.
3. The use of test-taking strategies can distort the outcomes of conventional item-based measures, but prior studies suggest that this may not be the case with VPAs.

4. VPAs can seamlessly incorporate features to minimize the importance of prior content knowledge and can track the extent to which a student utilizes these.

5. VPAs provide a more detailed record of student actions than do conventional item-based tests.

6. At scale, VPAs are as cost-effective and practical as other forms of digital assessment.

Our research is working to establish whether VPAs’ psychometric properties are sufficient to justify their use in high stakes testing, thereby realizing these advantages.

Implications for Policy

In this paper, we have discussed virtual performance assessments as a new model for re-conceptualizing the assessment of science inquiry. VPAs are based on over two decades of research on performance assessments and assessment design, cognition, technology, and videogame design. This type of assessment has considerable promise not only for measuring higher-order skills in science, but also for evaluating students’ progress on other sophisticated intellectual and psychosocial performances.

We offer the following recommendations for the practical, scalable implementation of VPAs as part of comprehensive state assessment systems:

- Include virtual performance assessments as part of comprehensive state assessment systems. Given the advantages of VPAs discussed above, the redesign of state assessment systems should move beyond using technology to digitize and automate conventional assessments.

- Ensure technological capacity. When investing in infrastructure for digital assessments, schools and districts should purchase machines with modern videocards capable of displaying detailed graphics and animations. The ability to render visually rich environments, such as virtual worlds, is important not only for VPAs, but also for instruction in general. As schools increasingly use digital content to aid learning, the devices and networks they purchase should be capable of delivering the full features of this material.

- Provide teachers and students with opportunities to use virtual performance assessments. In order for new assessments to be a valid measure of students’ knowledge, the technology cannot be a barrier to them demonstrating this knowledge. Students must have experience with and comfort using computers in order for their knowledge and skills to be assessed on computers. Similarly, professional development for teachers should include opportunities to use virtual performance assessments. Teachers understand the format of paper-and-pencil, item-based tests and how to gear instruction to that type of measure. Being able to explore and use VPAs is important for teachers, as they then will know what types of learning experiences students need in order to perform well on these assessments. These experiences are essential in shifting what is taught towards the knowledge, skills and abilities that most matter in 21st century STEM learning.

- Provide opportunities for key stakeholders to experience virtual performance assessments. Parents, school boards, and community members should have the opportunity to experience performing tasks in a virtual environment that provides feedback on their accomplishments. This will alleviate fears that students are being taught and assessed in ways that maximize “game-play” rather than key educational objectives. Further, since VPAs simulate authentic practices in settings similar to the real world, this type of measure has a “face validity” that highlights why it is a valuable complement to traditional tests.

- Provide professional development to teachers to foster the instruction that will lead to high performance on VPAs. Educational stakeholders should be informed of the curricular and professional development investments necessary to increase student outcomes on measures of sophisticated intellectual and psychosocial performances. Teachers will need to understand both the knowledge, skills and abilities that underlie that domain and instructional methods for helping stu-
students attain that knowledge. Curriculum selection must include materials that foster deep experiential learning of core authentic practices in the domain. The focus of instruction must move from broad, shallow coverage to focus on fewer core topics. These are key educational improvements central to preparing students for further schooling, work, and citizenship in the 21st century that, if taught well, will be evident by student outcomes on VPAs.

Conclusion

This is not a time to be conservative about implementing new forms of assessment, such as VPAs. It is abundantly clear that digitized versions of paper-and-pencil, item-based tests are insufficient to assess vital STEM skills like science inquiry. In fact, substantial evidence shows that these tests are undercutting students’ learning the key knowledge, skills and abilities required for careers that could help the United States to compete in the global, knowledge-based economy (NRC, 2006; PCAST, 2010; USDOE, 2010; NRC, 2011). Policymakers and state education leaders have a once-in-a-generation opportunity to fundamentally improve the way students’ science-related knowledge, skills and abilities are assessed. We hope they have the courage and foresight to rapidly implement innovations that can foster students’ mastery of science and ultimately, their success in college, career and life.

Acknowledgements

Our Virtual Performance Assessments research is funded by the Institute for Education Sciences, U. S. Department of Education, grant R305A080141. The investigators’ findings, interpretations, and conclusions do not constitute an official position of the U. S. Department of Education. The authors gratefully acknowledge the contributions of Drs. Jillianne Code, Nick Zap, Geordie Dukas, and Michael Mayrath, as well as the assistance of numerous students at the Harvard Graduate School of Education.
References


President’s Council of Advisors on Science and Technology (PCAST) (2010). Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for America’s future. Washington, DC: Office of Science and Technology Policy, Executive Office of the President.


## Appendix A

### Table 1. Extended ECD Framework for VPA design and development.

<table>
<thead>
<tr>
<th>Modified ECD framework</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Domain Analysis</td>
<td>Develop purpose for assessment. Compile research on the development of inquiry skills. Develop definition of competence. Develop knowledge, skills, and abilities (KSAs) we are measuring. Consult experts in the fields about our chosen definitions and definitions of inquiry and assessment objectives.</td>
</tr>
<tr>
<td>II. Domain Modeling</td>
<td>Use information from the domain analysis to establish relationships among proficiencies, tasks, and evidence. Explore different approaches and develop high-level sketches that are consistent with what students have learned about the domain so far. Develop narrative descriptions of proficiencies of inquiry, ways of getting observations that evidence proficiency, and ways of arranging situations in which students provide evidence of targeted proficiencies. Create graphic representations and schema to convey these complex relationships, and develop prototypes.</td>
</tr>
</tbody>
</table>
| III. Conceptual Assessment Framework:  | 1. Cognitive Model: Identify set of theory or beliefs about how students represent knowledge and develop competence in a subject.  
  2. Student model: What complex of knowledge, skills, or other abilities should be assessed?  
  3. Observations/Tasks: Identify kinds of tasks or situations (interactions) that will prompt students to say, do, or create something that demonstrates important knowledge, skills, and competencies.  
  4. Evidence: Identify behaviors and performances that reveal knowledge and skill identified in the student model. Identify and summarize evidence.  
  5. Interpretation: Develop a method for interpreting observations and evidence. |
| IV. Compilation:             | Develop purpose for assessment. Compile research on the development of inquiry skills. Develop definition of competence. Develop knowledge, skills, and abilities (KSAs) we are measuring. Consult experts in the fields about our chosen definitions and definitions of inquiry and assessment objectives. |
| IV. Four-Process Delivery Architecture: | Develop architecture and processes for implementing assessments. Develop back-end architecture that will capture and score student data. Develop prototype. Pilot. |
| VI. Refinement               | Refine assessment based on pilot data. Iterative cycle. |
3.1: Student gathers data that help explain or provide evidence to justify the claim being made. 3.2: Student determines which data from a specific investigation can be used as evidence to address an explanation. 3.3: Student distinguishes credible data from non-credible data in terms of quality. 3.4: Student is able to gather data after the experiment that will provide the evidence needed to prove or disprove whether the causal relationship was true. O VX n: Observational variables associated with each KSA. Score for each step can be either zero or one.