

**A System of Assessment for the Next Generation Science Standards in
California:
A Discussion Document**

Prepared by the Stanford NGSS Assessment Project Team (SNAP)

November, 2015.

Jonathan Osborne, Ray Pecheone, Helen Quinn,

Nicole Holthuis, Susan Schultz, Jill Wertheim & Paolo Martin

SCALE

Stanford Center for Assessment, Learning, & Equity

Stanford | GRADUATE SCHOOL OF
EDUCATION

This paper has been produced by the Stanford Next Generation Science Assessment Project that has been funded by the S.D Bechtel Jr. Foundation to undertake work to develop a model of a system of assessment for the Next Generation Science Standards (NGSS) that are currently being implemented in California. This paper provides a framework for the ideas that are guiding the project and our recommendations for the system that would support the implementation of NGSS in California

Executive Summary

California (CA) has adopted the Next Generation Science Standards (NGSS) that are very different from the prior Californian science standards. To address these new standards and meet federally mandated requirements for state testing, a new system of assessments will be needed. Such an *assessment system* must help teachers and schools understand the vision of science education embodied in the NGSS, and support assessment for learning as well as assessment for accountability. Furthermore, to realize the vision of the Next Generation Science Standards, all students must have access to high-quality learning experiences aligned to NGSS standards. This requires the teaching of science at all grades in elementary school and access to all sciences courses in high school.

What has changed?

The prior California science standards stressed chiefly what students should “know” or “understand”. They also included “Investigation and Experimentation Skills” standards that introduced some knowledge of scientific practices. Assessments consisted of multiple independent selected response (multiple choice) items that chiefly tested definitions and terminology, rather than the ability to engage in scientific practices or reasoning.

The NGSS differ from the previous science standards in two crucial ways. First, it requires students to engage in one or more *science and engineering practices* while applying *crosscutting concepts* as they utilize the *disciplinary core ideas* of the science that they are learning. Second, the goals of this ‘three-dimensional’ learning experience are defined not in terms of what students should know and understand but rather in terms of a set of expectations of what they should be able to do using their understanding of science, called ‘performance expectations’. Third, the NRC Framework recognizes that science learning is most effective when students revisit and deepen their understanding of core concepts as they progress across the grade levels, making connections between ideas rather than treating each course or unit as a separate and disconnected entity. Thus the assessments of these standards must not only test knowledge of the content of science, but also assess students’ ability to engage in one or more of the scientific practices and their ability to make connections across disciplines using the crosscutting concepts.

What should be the Goals of a System of Assessments and Assessment Supports for the Teaching and Learning of Science?

The example assessment system presented in this paper was developed to satisfy multiple goals:

1. To communicate and clarify the intent of the NGSS to drive teaching and learning in an appropriate direction as teachers read the intent of any new curriculum from the assessments used to measure its outcomes.

2. To support the teaching and learning of science at *all* levels—from early elementary through high school.
3. To address all three dimensions of the performance expectations of NGSS. This will require a variety of tasks including performance-based tasks. Further, the assessment system will need to support the development of teacher and administrator expertise in using such assessments by providing a curated bank of examples of curriculum-embedded performance tasks that can be used to assess student performance on all three NGSS dimensions, and by professional development to support their use.
4. To provide data on performance at the student, class, school, district and state level. In the exemplar model we offer data are generated not only by mandated state assessments but also by classroom-embedded assessment using the curated task bank.
5. Finally, to fulfill federally mandated requirements while prioritizing the preceding goals.

What are the Critical Elements of an Innovative System of Assessments and Assessment Supports?

To respond to these challenges, the state will need to develop a system of assessments that supports deeper levels of learning for all students. To achieve this goal our example system includes a mix of state-required, common mandated assessments using a variety of task types, coupled with a curated task bank that including short and extended performance assessment tasks and rubrics for use as measures of student learning. Central to the development of a new system of assessments will be a mutually shared responsibility for curriculum and assessment and related professional development between the California Department of Education (CDE), county offices of education, and local districts, schools and networks. Based on the California NGSS framework, and as required by federal mandates, 3 grade levels will be targeted for standardized state assessments. For the purpose of this paper, we are assuming that the federally-mandated testing (also know as ‘state testing’) will be at grades 5, 8, and 11. State supports for assessment in the other grade levels will need to be designed to support the teaching and learning of science, to collect data on student performance, and to provide an incentive to teach science in all grades.

Table 1 below illustrates our conception of an innovative and feasible system that includes two major elements: 1) required federally mandated testing at three grade levels that includes short performance tasks; and 2) support for periodic high quality assessment at all grades 1-11 through a moderated secure task bank and associated teacher training and support materials.

Table1: Proposed System of Assessment for California Science Assessment

Grade	Part 1: External Mandated Tests		Part 2: Periodic Classroom Assessments	
	Component A: Multi-item types	Component B: Performance Tasks	Component C: Stand-alone Performance Tasks	Component D: Curriculum Embedded Performance Tasks (CEPT)
	<ul style="list-style-type: none"> Variety of item types including selected and constructed response Computer-scored 	<ul style="list-style-type: none"> Two short performance tasks Scored by trained group of teachers 	<ul style="list-style-type: none"> Shorter Optional () State-endorsed Teacher-scored Use is reported and student performance is used locally to monitor student progress and to provide meaningful feedback to improve student learning 	<ul style="list-style-type: none"> Longer Some may be required Curated open source Task bank, state endorsed and monitored to and made available to all schools and districts in California Teacher-scored Use & scores are reported
1 st – 4 th			(x)	(x)
5 th	x	x	(x)	(x)
6 th – 7 th			(x)	(x)
8 th	x	x	(x)	(x)
9 th – 10 th			(x)	(x)
11 th	x	x	(x)	(x)
12 th				

The mandated elements of the state assessment assume computer-based testing, because it allows a richer range of computer-scored short tasks in the multi-task test. Moreover, it is essential for short performance tasks as it allows students to interact with simulations. Computer-based testing thus facilitates a wider range of task design with more authentic performance elements than can be achieved with pencil and paper. Both NAEP and PISA are moving, or have moved, to computer-based testing for science for these reasons and both of these tests provide examples of tasks that could be adapted to align with NGSS performance expectations.

Developing, adapting or endorsing a task bank for classroom periodic performance tasks will be a significant new addition to the system of assessment in California. Nevertheless, we recommend this element strongly for two reasons. First, it will not be possible to assess all the performance expectations of NGSS in a valid manner using solely externally mandated state tests as the NGSS require the assessment of a much broader range of competencies than the previous system. Second, exemplary and curated tasks in the bank will provide an important means to define and communicate what the learning outcomes of science should be, thus guiding and promoting good science teaching and assessment practices within the state.

Part 1. Federally Mandated Testing

The mandated portion of this possible assessment system would require students to complete two components, that is two testing sessions with different test structures. Component A would consist of a test using multiple item formats (selected response, constructed response, scenario-based, simulations aligned with, and integrated across, the three NGSS dimensions: Disciplinary Core Ideas, Cross Cutting Concepts and Science and Engineering Practices). This test would be offered on-line and would require approximately 60 to 90 minutes to complete.

Component B would require students to complete one or two on-line performance tasks. The available tasks would give broader coverage of all performance expectations for three grade levels up to and including the tested year. These tasks would include scaffolds to support learning needs (e.g., ELL, universal access). Testing time would vary by grade from 45 to 90 minutes. Student responses would be hand-scored by a trained group of teachers using common scoring rubrics. Because of the limited time, it is going to be impossible to test all students on all of the performance expectations. Hence, the tests will have to use some form of matrix sampling from which a reliable score for individuals could be calculated.

Part 2: Periodic Classroom Assessments

A state supported test bank would provide curated model tasks for in-class performance assessments for grades K-8 and for a variety of high school courses sequences. This task bank could be part of a broader national effort, and tasks could be developed not only by state-contracted developers but also by other science education coalitions and networks, or by individual districts and schools, provided they are accepted into the bank through the curation and revision process. These tasks would be designed to support deeper learning and to prepare all students for college and career success.

The test bank would include two types of tasks. Component C of our example system consists of stand-alone, short performance tasks together with appropriate scoring rubrics for each task and examples of how such rubrics are to be used. These tasks would be designed to be similar to the tasks for Component B of the external mandated testing to help inform teachers and students what to expect in the mandated test, and provide opportunities for formative assessment of student progress. Component D consists of longer curriculum-embedded performance tasks (CEPT) that would be embedded in the learning cycle of a unit. These would include curriculum and instructional resources to support the intended learning sequence and specify the individual student work products to be used for assessment. The task bank would also provide scoring rubrics and sample student work at each scoring level. The class time needed for these tasks with embedded assessment would be up to one to two weeks of instruction time, of which the assessment element of one or two lessons would be built into the curriculum across the unit of study.

Use of Component C tasks would be optional. Use of Component D tasks could either be optional or could be required in the tested grades. The state might require reporting of the use of any of these tasks

(i.e., which students undertook which tasks, with no recording of scores) as part of a system that monitors and encourages teaching of science in the untested grades. Districts, schools and individual teachers could use these tasks as formative or summative classroom assessments, and as models for tasks that they design.

Summary

A system that is limited to one or two end-of-year testing sessions at three grade levels cannot meet the primary goals of encouraging and supporting good science teaching across all grades. Conversely, a system that is based solely on in-class performance assessments cannot provide the reliable individual student and school-by-school student sub-group reporting that is required by federal mandates. Hence, we propose a mixed model with some elements of each.

This model suggests a large number of decisions that need to be made beyond the grade levels and standards to be tested to meet federal mandates. Each decision has a resulting set of costs and benefits to be evaluated. No matter what system is chosen, significant new investment in task development will be needed for all of its respective parts.

A System of Assessment for NGSS Science in California

1. The need for a new approach

Why new science tests?

California (CA) has adopted the Next Generation Science Standards (NGSS)—a set of science standards that are very different from the prior Californian science standards. The NGSS were developed from the vision for science education presented in *A Framework for K-12 Science Education* (NRC, 2012). To support that vision, the expected outcomes for students in NGSS are defined not by what students should know, but rather as a set of performance expectations – that is what students should be *able to do* with what they know. To address these new standards and meet the federally mandated requirements, a new system of assessments and assessment supports will be needed. Moreover, such an “assessment system”¹ must also help teachers and schools understand the innovative vision of science education embodied in the NGSS and implement it in their approaches to science teaching and learning. For instance, federal and state rules require standardized accountability measures. These are unlikely, however, to be maximally useful for formative purposes. But the accountability measures must include the sorts of tasks if they are to support and encourage the use of the formative and classroom materials. In other words, the system must seek to support assessment for learning as well as assessment for the sake of accountability.

Assessments are an important means of communicating the intent of any standards and operationalizing the content, practices and skills to be learned. Given the inherent ambiguity in any set of standards, teachers have a tendency to infer the specific meaning of the standards from the tasks that are to be used for assessment often designing their instruction to prepare students for that assessment. Thus, it is critical that the State assessment system be designed to convey and support the intended vision for science education in the NGSS.

Further, the NGSS are written from a position of “All standards, All students”, that is that a well-rounded science education should be available to all students, not just the few who may eventually become scientists or engineers. Today everyone needs to develop an understanding of science and the ability to engage in investigatory and design practices both for his or her role as citizens and for their future employability. For instance, the ability to ‘obtain, evaluate and communicate information’ is a competency that is essential in most forms of employment. Developing such capability, however, requires opportunities for all students to learn science, starting in early elementary grades and continuing through to high school. Yet, currently in many districts, the time allocated to learning and

¹ We use the term “assessment system” since, as we will show, any one test cannot meet these new goals particularly well

teaching science in the early elementary grades is one hour a week or less². Moreover, an emphasis on testing language arts and mathematics has resulted in reduced opportunities to learn science, particularly for students in high-needs schools and for language learners. Given the importance of early experiences of science for engaging and stimulating students' interest in the subject³, limiting the provision of science instruction negatively impacts students' subsequent opportunities to pursue and succeed in STEM coursework, which in turn restricts individuals' career choices. Ultimately, weak science educational opportunities have broader negative societal and economic impacts.

Likewise at the high school level, the state requires only two years of science for graduation. The likelihood that a student will take three and four years of science correlates highly with socio-economic status. For all students to have access to a learning experience that meets the standards will require an opportunity for the majority of students to take more than the minimum requirement of two science courses. Schools in CA are developing a variety of pathways through high school, including the course sequences that include science and engineering learning opportunities. Thus, a new assessment system will need to incentivize and support districts to consider the expansion of science in the high school curriculum rather than marginalizing science and suppressing access to high-quality programs for all students.

What has changed?

Previous CA science standards stressed chiefly what students should “know” or “understand”, privileging core knowledge focusing on recall of memorized facts. The “Investigation and Experimentation Skills” standards defined another class of knowledge, and the items designed to test these standards were designed to be “content neutral” in that they did not require any specific science knowledge. Consequently, the assessments used to measure performance on these standards chiefly tested definitions and terminology using multiple independent selected response items and, it can be argued, were not a test of students ability to engage in scientific reasoning.

The vision for science learning developed in *A Framework for K-12 Science Education* and underlying the NGSS differs from the previous science standards in two crucial ways. First, it introduces a notion of “three-dimensional” science learning, in which students engage in one or more *science and engineering practices* while applying *cross-cutting concepts* as they seek to understand phenomena or design systems that require them to apply and use the *disciplinary core ideas* of the science that they are learning. Thus, the NGSS introduces and defines a new set of eight science and engineering practices and a set of seven cross-cutting concepts (concepts with applicability across all science disciplines). Second, the NRC Framework recognizes that science learning is most effective when students revisit and deepen their understanding of core concepts as they progress across the grade levels, making

² Dorph, R., Shields, P., J., Tiffany-Morales, Hartry, A., & McCaffrey, T. (2011). High hopes– few opportunities: The status of elementary science education in California. Sacramento, CA: The Center for the Future of Teaching and Learning at WestEd.

³ Tai, Robert H., Qi Liu, Christine, Maltese, Adam V., & Fan, Xitao (2006). Planning Early for Careers in Science. *Science*, 312, 1143-1145.

connections between ideas rather than treating each course or unit as a separate and disconnected entity. To achieve these goals, the NGSS standards are written as performance expectations. Each of these includes a science and engineering practice that draws on a core disciplinary idea, and often also require the application of a cross-cutting concept. Consequently, the assessments aligned to these standards must test not only knowledge and understanding but also assess students' ability to engage in one or more eight scientific practices and their understanding of 7 cross-cutting concepts. Thus, ideally any one task (which may be a cluster of items) should test these three dimensions (or at least two of them), though not necessarily be able to separate the performance scores separately for each dimension.

What do these changes mean for Science Assessment?

Tasks designed to test such a multifaceted performance expectation will inevitably be longer than a single, selected response item, and will, therefore, take students more time to complete. In the classroom teachers will need to assess students' learning using extended, curriculum-embedded tasks that engage students in the NGSS practices and in using the cross-cutting concepts. A single task in which students engage in science and engineering practices to investigate and explain a phenomenon or design a solution to an engineering may need to extend over multiple lessons, with opportunities for assessment embedded in the task. Such tasks blur the boundaries between a learning activity and an assessment activity, being in part both. How then can the State assessment system support and incentivize teachers to use this approach for formative classroom assessments and for summative grading?

Given the "three-dimensional" approach to science learning called for by the NGSS, the use of performance tasks will be required, in some instances, to measure outcomes which are a combination of practices, concepts and ideas. By a performance task we mean a task where students must perform science or engineering practices to solve an open-ended problem. In addition, to address the three-dimensional nature of the NGSS, external tests will need to include technology-enhanced items, open-response and performance tasks, including simulations in order to measure the full range of the performance expectations defined by NGSS.

Some of these task types will require more time than traditional multiple-choice and short answer questions with which most teachers and students are most familiar. For example, short performance tasks may take 20 minutes to a full class period, and may form part of a formal, summative test. However, in any one test – typically of 60 minutes – it will be difficult to make any general statement about student competency across all the four domains of physical science, life science, earth and space science and engineering. Therefore, it is unlikely that the test would include multiple independent tasks addressing the same standard. This poses a design challenge, both for how to ensure sufficient reliability and generalizability of individual student scores, and how to decide what range of standards must be tested if the tests is to provide a valid measure of the effectiveness of a school's science program.

For instance, if such a test only has 5 tasks each consisting of multiple items the predictable complaint will be that for any one student, one or two or three of those item clusters happened to be "about"

something she or he did not like, was not interested in, or did not have experience with and, that if she or he had only been given one or two or three of the other item clusters, she or he would have done better. Of course, this is a complaint raised about any assessment (which is after all a sample from a domain of knowledge and skills), and in particular one that uses performance tasks.

One solution is to use extended measures of performance – what we choose to call ‘curriculum-embedded performance tasks’ that require students to engage over an extended period, that is (over multiple class periods), in one or more of the science or engineering practices, and to record or display the results of that work. A performance task does not necessarily require a hands-on investigation. For example, it could involve using a simulation to investigate a phenomenon, or analyzing and interpreting data collected by others. Our notion of a performance task is one where a student has some potential to choose how to address the problem, and where the scope of engagement in the practice is broader than a single element of the practice. In the extended curriculum-embedded tasks, students have opportunities to revise and refine their work, based on teacher or peer input. A major design question then will be how to support and incentivize teachers to use such tasks in the classroom and whether, and how to include any results of such classroom-based performance assessment in the scores recorded for state assessment purposes.

Finally, assessments are moving to computer-based platforms. From an educational perspective, computer-delivered assessments typically permit a much wider range of competencies to be tested, for example if the technology uses well-designed simulations or other computer-based tools allowing test-takers to demonstrate whether they can carry out investigations or to interpret data appropriately. However, as with all forms of summative testing the final test must resolve multiple constraints of testing time, cost, requirements for universal access, as well as the desired capabilities and tasks.

All of these new demands and the on-going emergence of new computer capabilities argue for a staged approach to design and implementation of a new forward-looking system of assessments and assessment supports. In the near term, such a system should be compatible with the technology and platforms already in use in California for Smarter Balanced Consortium (SBAC) testing of mathematics and language arts.

2. What should the Goals of a CA system of Assessment and Supports for Assessment in Science be?

Goal 1: Support the vision of 3D science teaching and learning

Any system of assessment must help to communicate and clarify the intent of the NGSS and drive teaching and learning in appropriate directions.

At all grade levels, the state or Districts will need to provide for the development of a range of assessments that support the teaching of learning and science and the attainment of the performance expectations of NGSS. Teachers will need to be supported by professional development

and shown how to use formative and summative assessments to adapt curriculum and instruction to meet the knowledge demands of NGSS.

Goal 2: Support the “All students, All standards” vision of NGSS at all grade levels

To monitor the implementation of the NGSS districts will need to develop measures/reporting of inputs not just outputs, for example opportunities to learn science, the resources devoted to science, and the patterns of science course taken by students. By tracking and collecting these data, each district will be better able to recognize and remedy inequities and deficiencies in their science offerings and teaching assignments.

Any system of assessment should support the teaching and learning of science at all levels. Of particular concern is the neglect of science teaching in the elementary grades if it is not tested. The state or Districts should monitor the quality of science teaching in early years for example by introducing locally developed, shared and vetted tasks and resources that measure the standards attained by the school, possibly at the beginning of 4th grade and which would provide formative feedback to the 4th and 5th grade teachers.

High school assessments should be based on the premise that all students have had the opportunity to learn the material in all disciplinary areas of the natural sciences (physical, life and earth) in a context that includes engagement in both science and engineering practices and application of crosscutting concepts. For most students this will require them to study more than the minimum required two years of science and the timing and nature of high school assessments should support this option.

Goal 3: Incentivize and support teachers to use effective formative and summative classroom assessment strategies

The system needs to support the development of teacher and administrator expertise in assessment by providing:

- a bank of examples of curriculum-embedded performance tasks that can be used to assess student thinking with scoring rubrics that exemplify how the results might be interpreted and used to inform further instruction.
- examples of approaches to summative assessment that support and reinforce three-dimensional learning of science – that is examples of the performance tasks and exemplar items to be used in the state assessments.
- professional development to show how tasks and items can measure students’ ability to meet a set of performance expectations that are significantly different from the knowledge and understanding previously tested.

Those responsible for assessment in a District will need to work with the professional development division, professional teacher networks (e.g., CORE), and/or service centers (County Offices) to support teachers and districts to learn to use these tools to support and improve the quality of their teaching.

Goal 4: Help to provide needed data at the student, class, school, district and state level

Data from assessments need to be provided as rapidly as possible and in an accessible manner to teachers and administrators (school and district level) if they are to be useful in informing instructional decisions and professional development.

Goal 5: Meet Federally Mandated Requirements

While it is important that the system meet the federal legal requirements, either in their current or new incarnation, it is important that any assessment system is not restricted by any narrow interpretation of what is important. The goal should be to develop a quality system of assessment which supports students in their learning of science and teachers to improve their pedagogy which, meeting the federally mandated requirement, does not simply to focus on the minimum needed to meet federal requirements.

Any system of assessment should be consistent with the NGSS standards, accessible to all students and should not overly restrict options for the sequences of middle and high school science course.

Parents, students, and teachers need to understand the system and its rationale. Therefore, the decisions on what grades and standards will be tested for federally mandated purposes, the goals and purposes of the assessment, and how the scores of these tests will be used should be transparent and made public through as many communication channels as possible. For districts, such information is important to allow them to make scope and sequence decisions about the course options and curriculum. Ideally these new course sequences need to begin to be implemented at least two years before any testing is used for accountability purposes.

3. What are the recent advances in science assessment capabilities?

A major development in assessment in the past decade has been the move to computerization. At the moment, computerization offers some potential advantages over conventional methods of assessment. A greater range of item types can be developed to accommodate the language and special needs of a diverse set of students. In addition, computer-delivered (both adaptive and non-adaptive) tests allow new task and item types that have no paper and pencil equivalents and can be, in many cases, machine scored. For instance, students can be asked to reorder a list into the correct sequence, drag and drop items from a list to the appropriate gap in some written text, or run a simulation of a phenomenon to collect data. Simulations, for instance, allow a variety of complex tasks, some of which are simulated performance tasks. Several examples of these were trialed in the NAEP 2009 assessment⁴. At the moment, many of the simulation-based tasks focus on the design of experiments to test hypotheses and assess students' ability to identify the relevant variables, control the appropriate variables, collect an appropriate range of data, and interpret the results. While this is undoubtedly one domain of science worth testing, other areas of reasoning such as the ability to develop models, categorize and classify,

⁴ See http://www.nationsreportcard.gov/science_2009/ict_summary.aspx

undertake probabilistic and statistical thinking, or construct arguments of an inferential nature are not so readily tested. Further work needs to be done if such simulations are to test the full range of scientific thinking to be found in the NGSS. However, in general, the capability of technology to offer a greater range of forms of testing is a significant change in the assessment landscape, and one that is particularly important in the area of science.

A second advantage of computerization is that it is capable of offering testing that is adaptive to the skills and understandings demonstrated by a student. The expected advantage of such a system is twofold: a) items are matched to the student's level of understanding which provides a test with greater discrimination and a better measure of student understanding – in short, a more valid test; b) the weak student is able to answer a higher proportion of the questions correctly, and hence feel less incompetent and more encouraged by their testing experience. More able students are presented with questions that are better matched to their level of knowledge and understanding so that the assessment is more challenging and possibly more engaging. Such testing could be designed to offer diagnosis of incorrect student responses to complex tasks by presenting related but simpler tasks. However, while adaptive testing for mathematics has been implemented in large-scale systems, the problem for science is much harder as, because of the wide range of topics, it is difficult to use responses on one topic to predict what items to offer on a different science topic even within a given grade level. Hence, as far as we know, no large-scale systems are using adaptive testing yet and, based on a thorough discussion of testing options with our technical advisory committee (TAC), we are not recommending development of adaptive testing for science at this time.

A third advantage of computerization, is that machine scoring⁵ enables the teacher and the student to be provided with instant feedback about student performance supporting the use of assessment for formative goals. As Hattie argues in his book *Visible Learning*⁵ – which is a review of major contributions to effective teaching – “when teachers seek, or at least are open to, feedback from students as to what students know, what they understand, where they make errors, when they have misconceptions, when they are not engaged, then teaching and learning can be synchronized and powerful. Feedback to teachers helps make learning visible.” The problem for most teachers is simply a data problem – with 30 students in a class, hand grading takes time. As a consequence, much of the data loses its salience because the teacher and the class have moved on to the next topic by the time it is completed. In contrast, instant feedback provides the teacher with the information necessary to make informed pedagogic decisions about the nature of student learning difficulties and where best to apportion their help and assistance. For the student, well-designed feedback can be an important aid to learning and can stimulate reflection on progress. Furthermore, computer-based tasks can allow students to ask for feedback at various stages of their work, and to use that feedback to help them decide their next steps or to revise and improve their work. Another advantage of computer-based testing is that it enhances accessibility for all students, particularly for those with disabilities.

⁵ Hattie, J. (2008). *Visible Learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Of course, computerization of science assessment comes with challenges as well as opportunities. Designers must ensure that the technology does not become a barrier to the measurement and that access to better and more expensive technology does not broaden group differences for reasons other than true domain differences. While they offer the opportunity to be engaging, computer interfaces can also be confusing, particularly if one cannot assume all students “practice” on the specific interface to be used in the assessment. These cautions notwithstanding, computerization offers the real possibility of revolutionary improvements in science instruction and assessment.

Another major imperative driving tests in science is the desire to go beyond testing recall. For instance, the OECD PISA science test administered globally to representative samples of students from over 60 countries seeks to test three ‘competencies’. These are the ability to: (1) explain phenomena scientifically, (2) evaluate and design scientific inquiry, and (3) interpret data and evidence scientifically⁶. These competencies are seen as not just requiring content knowledge but also a knowledge of the procedures that science uses to establish its claims to know and the epistemic constructs and criteria used to make scientific judgments. In addition, the PISA tests categorize items by three levels of cognitive demand – low, medium and high and items that ask only for recall are classified as low cognitive demand items. A select few items that we have obtained from our initial review that go some way to meeting these requirements and those of NGSS are described in a separate report (Wertheim et al., in progress).

Likewise the international TIMSS tests⁷ seeks to test students’ knowledge of science (40%), the ability to apply such knowledge (40%) and the ability to reason scientifically (20%). ‘Applying’ is seen as the ability to use scientific knowledge to compare/contrast/classify, relate observations to underlying science concepts, use models, interpret information and explain. ‘Reasoning’ is seen as the ability to analyze, synthesize, formulate questions, design investigations, evaluate, draw conclusions, generalize and justify. Similarly, the NAEP tests seek to test not only students’ knowledge but also students’ ability to identify science principles, use science principles, or engage in scientific inquiry and technological design. The primary driver of these more complex and sophisticated tests is the growing recognition that education must do more to develop the higher order skills necessary for employment in today’s society.

Another body of knowledge with the potential to inform the quality of assessment is the growing quantity of work that has been undertaken on learning progressions. Learning progressions are hypothesized progressions of student understanding which are tested empirically and iteratively to determine ‘how students’ understanding of, and ability to use core scientific concepts and explanations and related scientific practices grow and become more sophisticated over time with appropriate

⁶ OECD. (2012). The PISA 2015 Assessment Framework: Key competencies in reading, mathematics and science. <http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm>

⁷ Jones, L. R., Wheeler, G., & Centurino, Victoria A.S. (2015). *TIMSS 2015 Assessment Frameworks*. Boston: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston.

instruction⁸. As such, the progression they map can guide the design and structure of good assessments. However, in general, there is, as yet, little empirical evidence about how a students' knowledge of specific scientific domains or specific scientific practices develops over time. Consequently, any item writer makes an implicit judgment about the suitability of the item and what it assesses – essentially a process of professional judgment and trial and error. Thus, the work on learning progressions can inform assessment in two ways. First it suggests aspects of concepts that are likely harder for students to understand or to practice. Second, such work has established a body of items that have been used to test students extensively and are accompanied by scoring guides. While such learning progressions do not exist for all disciplinary core ideas or all of the 8 practices in NGSS, where they do exist, they provide empirical evidence of the range of performance students might be expected to achieve. Hence, they provide an important benchmark for framing and designing assessment tasks and eventually for building assessment tasks that test the full range of ability.

4. What decisions are needed prior to the start of the design of new tests?

There are three distinct areas of decisions that need to be made prior to the design of any new tests. These are decisions about the overall system of assessment; a set of choices about meeting the federally mandated requirements; and a set of decisions about what other testing might be beneficial to teaching and learning in California. In this section, we simply present a series of questions that need to be considered in developing the tests that assess the California NGSS.

4.1 Overall systems design decisions

1. First, the state must decide on what are the priorities and goals of the state for a new system of assessment and support for the NGSS standards. Who, at all levels of the system, needs to be involved to answer this question (i.e., parents, teachers, schools, districts and state policy-makers)? How can users of the system inform policy and be engaged in a dialogue about the system with major stakeholders such as school districts, legislators, teachers, community groups, and the California Science Teachers Association and parents?
2. What are the constraints on this design process such as deadlines, testing time, costs, technical requirements to yield valid and reliable scores, and legal obligations that must be met as the design is developed and elaborated?
3. What data must the system generate, and for whom (individual scores, school, and district-level accountability)? How will the data be stored, maintained and accessed, and by whom?
4. What are the stakes (and consequences) for students and schools associated with scores in this system? In what other ways may the data be used?
5. What are the challenges that can interfere with the successful implementation of such a system? Likewise, what opportunities does the new system present to address the limitations

⁸ Corcoran, Tom., Mosher, Frederic A., & Rogat, Aaron. (2009). Learning Progressions in Science: An Evidence-based Approach to Reform. Philadelphia, PA: Consortium for Policy Research in Education.

of the previous system such as the narrow range of knowledge and skills tested by the previous system?

6. Should a pencil and paper version be an option, and if so, under what conditions and for how long? If so, how will it be similar to/different from the computer-based version? If so, must it be identical to the computer-based version? If so, it will be impossible to take advantage of many of the facilities offered by the new platform. Alternatively, should the pencil and paper version be a limited subset of the main test?

4.2 Decisions about federally mandated tests

7. What grades will be tested for the federally mandated science tests?
8. What standards will be eligible to be tested at those grades? Will it just be the standards in the grade tested or all of the standards since the last test?
9. What flexibility does the state have to use matrix-sampling methodologies at these grades?
10. What is the desired ultimate form of the test? Based on the development of an assessment system using multiple measures, what proportion of items should be multiple-choice items, constructed response items, technology-enhanced items (including simulations) and performance based items that require hand scoring?
11. How can the state build a system that is flexible and adaptable with improvements in technology so that eventually it can use adaptive testing, incorporate process interaction with simulations, or add capabilities, e.g., for students to draw graphs and more?
12. Should data from the assessment be combined with scores from performance tasks done over multiple days in classrooms for accountability purposes?
13. What role should local development, if any, play in the design of a state assessment system?
14. What type of alternative assessment will be developed for students with severe cognitive disabilities and will this assessment measure the full range of standards?

4.3 Decisions about “other grades” tests and other elements of the system

15. What assessments and systems of support for assessment will be provided by the State at grades other than the mandated years? What resources can/will the state provide for this work? What is the role of County Offices, Districts and Science Education coalitions in developing such a system?
16. How can classroom assessment (both formative and summative) that improves student learning best be encouraged and supported by the state system?
17. How can performance tasks requiring hand scoring be introduced, scored reliably, and contribute to individual scores?

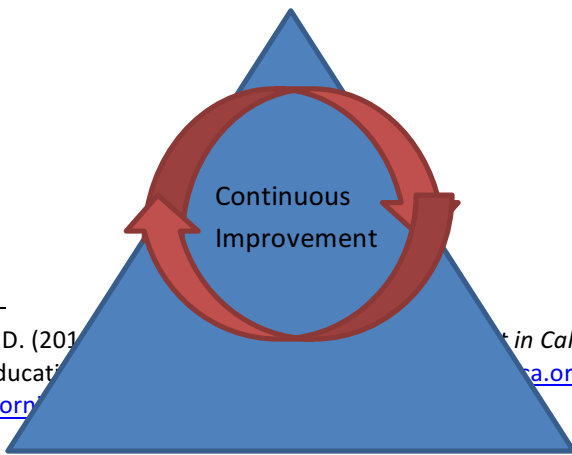
5. What are the critical elements of an innovative system of assessments and assessment supports?

The adoption of new standards has become a catalyst for raising expectations for student performance with the potential to impact teaching, schools and classrooms nationwide. In California, the Local Control Funding Formula is designed to put in place, at all levels of the educational system, support structures that promotes system-wide learning and which can transform curriculum, instruction and assessment to better ensure student readiness for college and/or career. To respond to these challenges, and support the proposed changes in the design of new accountability systems in California, our view is that the state will need to reimagine and reinvent accountability to drive the change toward a more equitable system of assessment that supports deeper levels of learning for all students. To achieve this goal a mix of state-required, common mandated assessments coupled with locally developed assessments and measures of student growth are needed.

Any new system should be grounded in the concept of the reciprocal accountability – a concept developed in a recent paper by Linda Darling-Hammond and David Plank⁹. In such a system, every actor in the system must be responsible for the aspects of the educational quality that it controls. On the one hand, the state has made three fundamental commitments: (1) to pursue *meaningful learning* for all students; (2) to give schools and districts the *resources* and flexibility they need to serve their communities effectively; and (3) to provide *professional learning* and support for teachers and administrators at all the levels of the system. As shown in Figure 1, these constitute three pillars of a new accountability system that is designed to support continuous improvement:

- Accountability for meaningful learning (state plus local);
- Accountability for adequate and intelligent resource allocation (local reporting);
- Accountability for professional competence and capacity (local reporting).

Meaningful
Learning



Continuous
Improvement

⁹ Darling-Hammond, L. & Plank, D. (2014). *Reciprocal Accountability: A Policy Approach to Support Continuous Improvement in California's education system.* <https://www.cde.ca.gov/publications/supporting-form>

Figure 1: Key Elements of an Accountability System

In practical terms, the next generation of mandated assessments will incorporate innovative, selected response items that require deeper learning and the application of knowledge. They will include some number of constructed response, computer enhanced, selected response and performance tasks that, at present, require hand scoring. Items will often be grouped in clusters to form a task built around a single performance expectation. The clusters will be designed to require students to demonstrate core knowledge, scientific practices, and cross-cutting concepts (e.g., patterns, cause and effect, structure and function...) although not necessarily to generate separate scores for the three dimensions. In addition locally reported results should include those from more open-ended, curriculum-embedded, performance assessments including one or more of longer performance tasks, projects, portfolios, games, or extended use of simulations. Central to the development of a new system of assessment will be a mutually shared responsibility for curriculum and assessment between the California Department of Education (CDE) and local districts, schools and networks. Based on the California NGSS framework, and as required by federal mandates, 3 grade levels will be targeted for on-demand standardized external assessment. State supports for assessment in the grade levels other than those with required external testing will need to be designed to support, deepen and extend what it means to be proficient in science by focusing on three interlocking dimensions: Practice, Crosscutting Concepts, and Disciplinary Core Ideas. For the purpose of this paper we are assuming that the federally-mandated testing will be at 5, 8 and 11. (Note that the reasons for testing at grade 11 rather than grade 10 are discussed below though an alternative option is to leave the choice of grade 10 or 11 testing to the school or individual student).

Designing such a system requires many high-level decisions, and we do not presume to suggest we can make the right choices for the State. However, it is helpful to approach these decisions with a “straw man” example to help raise the relevant issues. Hence, we have sought to develop such an example, with input from our advisory committee to refine our initial ideas. In this example of an innovative and feasible system, we propose a system of assessments that includes two major elements: 1) required federally mandated testing at three grade levels (e.g., 5,8,11); and 2) support for periodic high-quality assessment at all grades K-11 through a moderated and curated secure task bank and associated teacher training and support materials. Requirements for use and reporting of data from the use of these tasks will be discussed below.

Table 1 below outlines this “straw man” system. This design is based on review and input from participants at our advisory panel meeting, March 2015. It also responds to the demand for sample items expressed by many groups at the stakeholder input meetings held by the state in the past year.

The mandated elements all assume computer-based testing, because it allows a richer range of computer-scored short tasks in the multi-task test and is essential for the short performance tasks as it allows students to interact with data or simulations. Computer-based testing thus facilitates a wider range of task design with more realistic performance elements that can be achieved with pencil and paper. Both NAEP and PISA are moving or have moved to computer-based testing for science for these reasons, and both of these tests provide some examples of tasks that could be adapted to align with NGSS performance expectations.

Table1: Proposed System of Assessment for California Science Assessment

Grade	Part 1: External Mandated Tests		Part 2: Periodic Classroom Assessments	
	Component A: Multi-item types	Component B: Performance Tasks	Component C: Stand-alone Performance Tasks	Component D: Curriculum Embedded Performance Tasks (CEPT)
	<ul style="list-style-type: none"> Variety of item types including selected and constructed response Computer scored May contain some matrix assigned elements 	<ul style="list-style-type: none"> Two short performance tasks Scored by trained group of teachers Matrix assigned 	<ul style="list-style-type: none"> Shorter Optional () State-developed Teacher scored Use is reported, scores are not reported to state but may be used by districts 	<ul style="list-style-type: none"> Longer Three options (see below) Task bank, state curated and controlled (eventually includes consortium/district developed tasks as well as state developed) Teacher scored Use & scores are reported to State
3			(x)	(x)
4			(x)	(x)
5	X	X	(x)	(x)
6			(x)	(x)
7			(x)	(x)
8	X	X	(x)	(x)
9			(x)	(x)
10			X	X
11	X	X	X	X
12				

We recognize that the task bank for classroom periodic performance tasks is a significant new addition to the system of assessment in California. Nevertheless, we recommend this element strongly for two reasons. First, it will not be possible to assess all the performance expectations of NGSS in a valid manner using solely external mandated tests as the NGSS require the assessment of a much broader range of competencies than the previous system. Such tasks will provide an important means to define and communicate what the learning outcomes of science should be, thus guiding and promoting good science teaching and assessment practices within the state. Additionally, the state could adopt or adapt an existing Resource Bank that is currently under development in science and other content areas nationally. For example, the Innovative Lab Network (ILN) task/resource bank is sponsored and developed by CCSSO and SCALE at Stanford University and California as a founding member of ILN could help build on this Resource Bank initiative to support NGSS system of assessment in California. At the current time, however, there are few NGSS-aligned science tasks in this task bank.

Part 1. External Mandated Testing

Key Features

The external mandated testing portion of this possible assessment system would require students to complete two components, with different test structure. Component A consists of a test that would utilize multiple item formats (selected response, constructed response, scenario-based, simulations aligned with, and integrated across, the 3 NGSS dimensions: Disciplinary Core Ideas, Cross Cutting Concepts and Science and Engineering Practices). These tests would have the following features:

- a) Focus on core skills and abilities within NGSS domains.
- b) Use universal testing and not adaptive (some items or item clusters may be matrix assigned to increase the coverage of the full range of performance expectations across three grade levels within the test.
- c) Include a mix of task types that are chiefly computer scorable, with possibly a few hand scored tasks.
- d) Assess capabilities beyond memorization and recall, with a focus on the application of knowledge (including crosscutting concepts) and the use of science practices.
- e) Be offered on-line and provide flexible use of scaffolds to support specific assessment needs (e.g., ELL, SWD)
- f) Be approximately 50-90 minutes.

Test Component B would require students to complete one or two on-line performance tasks. The available tasks would give broader coverage of all PE's for three grade levels up to and including the tested year. These tasks would:

- a) Use matrix sampling design that would enable estimates of school-level based achievement across the science standards and that also allows adjustments for comparability in calculating individual student performance scores.
- b) Include scaffolds to support learning needs (e.g., ELL, universal access).
- c) Allow testing time to vary by grade, e.g., 45 minutes (grade 5) 60 minutes (grade 8) and 90 minutes (grade 11).
- d) Require tests to be hand-scored by a trained group of teachers (as with AP tests) using common scoring rubrics (possibly eventual Artificial Intelligence scoring of some tasks).
- e) Include the use of simulations and/or data analysis tools such as manipulable spreadsheets or graphing capability.

Part 2: Periodic Classroom Assessments

A state supported task bank (test bank) would provide model tasks for in-class performance assessments at all grade levels 3-11. Support would require both advocacy and approval of a common platform, assurance of its accessibility to all schools, and the monitoring of a sufficient and effective

system of curation for alignment, quality, and developmental appropriateness. However, the state need not necessarily be the sole funder or provider of the tasks and the task bank collection. The task bank would include two types of tasks, short performance tasks similar to the on-demand performance tasks described above (Component C), and a set of longer (multiple class period) learning tasks with embedded assessment elements known as curriculum embedded performance tasks (CEPT) (Component D). Ancillary materials would include teacher guidance for use of these tasks including training modules and scoring rubrics for the assessment elements of the task.

The role of the state would be to:

- 1) Support the system that is providing open access to the bank for all educators in California.
- 2) Work or collaborate with other states and science education coalitions and testing experts to develop an initial set of items/tasks populating the test bank;
- 3) Specify required (if any) and optional uses for the task bank and encourage its use;
- 4) Ensure that there is a system for developing, curating and testing new tasks for addition to the test bank, including CEPTs developed by teacher groups or curriculum developers; and
- 5) Support county offices and other professional development providers to provide professional development on the use and scoring of such tasks.

Component C consists of stand-alone, short performance tasks. Our view is that the use of the short performance tasks should be optional. They will be similar to the tasks for Component B of the external mandated testing to help inform teachers and students what to expect in the mandated test and provide opportunities for formative assessment of student progress toward the desired performance level. The state could require reporting of the use of these tasks (which students undertook which tasks with no recording of scores) as part of a system that monitors and encourages teaching of science in the untested grades, with the record of the use of such tasks becoming part of the school level accountability system. Districts, schools and individual teachers could use these tasks as formative or summative classroom assessments, and as models for tasks that they design. Scoring rubrics (one for each task) will be provided across disciplines.

The Key Features of short performance tasks used periodically by teacher choice would be that:

- (a) Tasks would be designed to address all 3 NGSS dimensions but not to score them separately.
- (b) They capitalize, in part, on released items from on-demand testing.
- (c) Scoring rubrics and student work samples would be provided in the task bank along with each task.
- (d) They enable teacher choice.
- (e) They would include supports for teachers to understand the scoring rubrics to enable them to use them effectively and consistently with group moderation to develop common scoring standards across a grade level within a district

- (f) They include scaffolds that support learning needs (ELL)
- (g) The outcomes of student performance would be recorded and archived within the district and used for formative evaluation (i.e. to inform instruction) as well as to monitor student progress within and across grade levels
- (h) The approximate time to complete such a task would be 20-30 minutes

Curriculum Embedded Performance Tasks (CEPT)

Component D consists of longer performance tasks that would be embedded in the learning cycle of a unit. These CEPT tasks will be included as a component of the on-line curated task bank discussed above and provide additional resources including a sequence of the lessons in which these tasks are embedded and instructional tools and resources to support their use. Key features of the task bank are that:

- (a) The bank would include curriculum and instructional resources needed to support these tasks, specification of individual student products required for scoring, scoring rubrics and sample student work at each scoring level. The tasks could be populated from local development, research centers and networks (e.g., CORE, ConnectEd, SCALE, CRESST).
- (b) All resources would be peer reviewed and curated through a specified testing and improvement cycle before inclusion for general use.
- (c) All tasks would be tightly aligned to constructs also tested in the federally mandated assessment.
- (d) All tasks would include scaffolds to ensure that tasks are designed to support diverse learning needs with specific attention to the needs of special education and ELL students.
- (e) Common rubrics and formats would be provided to score tasks.
- (f) Scoring modules and on-line resources for teacher training to implement and score the tasks would be developed and made available on a website. County offices, research centers, and districts would be encouraged to support such training, including opportunities for the moderation of scores among a targeted sample of teachers.
- (g) The bank would be supported by a system of support for professional learning to build expertise with assessment at scale amongst teachers (e.g., asynchronous system of in-person and online learning (MOOC's).
- (h) The class time needed for these tasks with embedded assessment would be up to one to two weeks of instruction time, of which the assessment element of would be embedded in class instruction (equivalent to approximately 1 to 2 class periods).
- (j) There would be a system of recognition and incentives for organizations, county offices, science education coalitions, research centers and individual schools and/or teachers who contribute to the resource bank.
- (k) Research to examine the relationship between performance on the CEPT assessments and the on-demand assessment as well as other test metrics (e.g., ACT, SAT) would be helpful to build a deeper understanding of their impact on learning.

Note: California is a founding member in the development of a curated task bank developed by CCSSO and SCALE at Stanford University as part of the Innovative Lab Network (ILN) consortium that includes 18 states. California could adopt or adapt the ILN open sourced resource bank to address the needs of component C and D discussed above. This approach could optimize

existing resources and avoid duplicating efforts on a parallel development of a NGSS task bank. At present, this bank does not contain a significant collection of NGSS aligned tasks.

Three options are possible with these task banks.

Option 1: All such tasks would be voluntary and whether to use them would be a school or district decision. There are two arguments for their use. First, is that they would provide important feedback to teachers about the level of competence their students have achieved and be a better preparation for the tests required by the mandated tests—Component A and B. Second, such tasks would encourage a deeper learning of the scientific content and skills preparing students better for college and career success.

Option 2: These tasks would be required in the three tested years. Typically, students would undertake one or two tasks that would be graded by the teacher who should have participated in a program of scorer training. Scores on these tasks could count for up to 20% of the final external assessment score. Teacher scoring could be validated by random audits of teacher scores and/or by checking the range of teacher-assigned scores against the range attained in other parts of the federally mandated tests.

Option 3. In each year of grades K-8, and in each high school science course, each student would be required to complete 2 (1 per semester) at each grade level. Teachers (or the district) will choose which tasks to use from a curated bank of tasks. Scoring will be done by the teacher using a rubric provided that would be accompanied by examples of student work at each level. These scores could then, if desired, be reported to the state system and could be accumulated across a three-year grade span and used as a small element in overall student scores (10-20%) on mandated testing. Scores in the K-2 grades would be used only for internal district purposes. Again, teacher scores could be validated by comparing the range of student scores awarded by the teacher with those obtained on the federally mandated test. Such a system would provide an incentive to teach science at all grade levels, as well as reinforcing the message that students must engage in such tasks to achieve competence with the NGSS performance expectations.

Districts could keep records of student work on these CEPT tasks for internal progress monitoring and schools may be asked to provide samples of student work at each scoring level for review by external monitors as a part of a school and district-level accountability system. These monitors could be teachers from other districts who have been trained for this work. Districts and county offices will be encouraged to provide professional development to teachers on how to use and score such tasks, and the state would provide on-line resources to support such training and/or individual teacher learning.

6. What are the lessons to be learned from past assessment efforts in California and elsewhere?

Finally, we would point to the fact that there is much to learn from previous attempts to develop more ambitious assessment systems in California and elsewhere. SCALE conducted a retrospective research study of a range of performance assessment initiatives that began in the 1990s, tapping into the expertise of those with a deep reservoir of experience from these earlier efforts to integrate the use of performance assessments into large-scale assessment programs in the United States¹⁰. The study documents and synthesizes the common political, technical, and practical issues related to these earlier efforts with the goal of answering three specific questions:

- What were the conditions that helped sustain some of the programs?
- What were the challenges that led to their discontinuation?
- What are some lessons learned that might help inform current assessment initiatives that seek to integrate performance assessment into large-scale student assessment programs?

Through an extended literature search of available research papers, solicitation of internal reports from the administrators of the performance assessment systems, and interviews of key stakeholders, we gathered key information about: 1) the design of the assessment programs, including evidence about the technical quality of those assessments (reliability and validity); 2) the goals and policy framework of the performance assessments; and 3) the implementation of the assessment programs, including the cost of implementation. The findings from this work are presented in Appendix B. Emerging from this review are the following requirements for any system of assessment.

1. The use of good task design practices to construct assessments that meet intended purposes and meet standards of technical quality, using a mix of short response (selected or constructed), and both simulation-based and classroom-based performance tasks.
2. The inclusion of classroom based performance tasks as part of the assessment system. These tasks should be curriculum-embedded, and produce well-defined student work products to be scored following a well-developed scoring rubric. Along with the task and its instructional context, both the required products and the scoring rubric should be communicated to teachers administering the tasks in their classroom. There are two options for scoring such performance tasks. One is train and pay a team to do the scoring, most likely using scorers drawn from known expert teachers. This team would then score student work submitted by the classroom teachers. Alternately, the costs of hand scoring class-room-based performance tasks can be minimized by involving teachers in scoring their own students work. This requires a systematic professional development process before it is instituted, and an ongoing moderation and oversight process. Indeed, such a system has been seen to have significant professional development benefits, leading to improved teaching and learning,

¹⁰ Wei, R.C., Pecheone, R.L., & Wilczak, K.L. (2014). *Performance assessment 2.0: Lessons from large-scale policy and practice*. Stanford Center for Assessment, Learning, and Equity

where it has been well implemented. Such scores can be sufficiently reliable for use as part of a school and district-level accountability, but not at student-level or teacher level.

3. The provision of a curated resource bank of high-quality NGSS-aligned performance tasks suitable for use as formative and/or unit summative assessment tasks. This task bank should be accessible to teachers to support powerful instruction and assessment practices. The resource bank should include rubrics for scoring the tasks, and models for task and scoring rubric development to help teachers in developing further such tasks.
4. The support of teachers through professional development to use this resource as part of a coherent system of embedded assessments, curricula, and instructional supports.
5. Minimizing the cost of developing performance assessment tasks through economies of scale and cross-state collaboration.
6. Engaging with stakeholders more actively, and developing the capacity of educational leaders and policymakers to deeply understand and champion research-based reforms in assessment
7. Engaging with the public more actively, and provide timely, accessible information about the new assessment systems and the NGSS.

7. Final Issues for Consideration

We conclude in this section by pointing to a few salient issues that also needed to be considered in building an assessment system that is fit for purpose.

- a) It is very important that the development of an assessment system for the Next Generation Science Standards in California be seen as work in progress with a staged approach to change. New assessments to be introduced in 2016-2018 should be seen as an “interim” step towards a new system of assessments and assessment supports, because the work required to implement the full new system will need a longer time horizon.
- b) The performance expectations of NGSS can only be fully met and adequately tested if there are some performance assessments provided as a component of all phases of education i.e. elementary, middle school, and high school. Ultimately, the system should permit such assessments to “count” in scoring student proficiencies and school quality for federally mandated purposes.
- c) Grades other than those in which the externally mandated tests occur require assessments that support student learning and which enable schools to monitor student progress. We suggest that the state provide assessment resources for these years but limit required use of them. Districts should be encouraged to use sampling methodologies and tasks from the state resource bank to gain oversight of what is being taught in these grades and to include some information on science learning outcomes in their reporting. If any additional required assessment is to be introduced, we suggest a matrix-sampled assessment at the beginning of grade 4, with school-level results rapidly returned to teachers. This could serve the goal of providing formative information to upper elementary teachers and encouraging the teaching of science in all elementary grades. This

investment in early-grades could potentially have a more beneficial effect on science learning across the state than, for example, new end-of-course exams at the high school level.

- d) The task development and review (certification or curation) process should include a cycle of review by disciplinary experts who check for scientific accuracy to strengthen task validity. Teachers who have experience at the targeted grade level also need to be part of the design and review team along with assessment experts.

8. Concluding remarks

The example we develop above is intended as a “straw man” to illustrate the opportunities and the challenges in developing a state science assessment system that meets the goals discussed above. While there are many possible designs, we recognize that there are also multiple design constraints. We are convinced that a system that is limited to one or two end-of-year testing sessions at three grade levels cannot meet the primary goals of encouraging and supporting good science teaching and learning across all grades. Conversely, a system that is based solely on in-class performance assessments cannot provide the reliable individual student and school-by-school student subgroup reporting that is required by federal mandates. Hence, we propose a mixed model with some elements of each. The details of our model were developed by a group process at our advisory panel meeting. What they indicate is that there are a large number of decisions to be made beyond the grade levels and standards to be tested to meet federal mandates. Each decision has a resulting set of costs and benefits to be evaluated. No matter what system is chosen, significant new investment in task development will be needed for all its parts. One way development costs can be reduced is by partnering with other states to develop a shared system. The impacts of such a system should also be tracked over time, with subsequent adjustments to the system.

Experience with curriculum-embedded performance tasks in science not only in the work of the SCALE group with various consortia of schools, but also in other states, and in other countries where such an approach is used such as Australia and New Zealand, has shown that using such tasks, along with providing assessment information, provides a more effective learning experience. Such tasks not only support learning for the students who participate but also provide a professional learning experience for teachers who use and score the tasks. Teachers who participate in designing and testing the tasks learn even more. Thus the strongest advice of this report is that California should find a way to support a system that includes such tasks for students at all grade levels and to support the capacity-building for teachers across the state to use, score and eventually to participate in teams to design such tasks.

Any new system will require a communication network to inform parents, teachers, schools and districts what is coming, why the choices were made, and what to expect in the early years of implementation. It will require staged implementation including a period of refining and adjustment, with low stakes for students, teachers and schools. In summary, these are the opportunities and challenges afforded by such a system.

Opportunities

- (a) The promotion of teacher and school participation in design, development and implementation of performance assessments.
- (b) Support for local collaboration and engagement in developing and using assessments.
- (c) Building assessment capacity and knowledge through direct involvement of teachers (e.g., task development and teacher scoring) with moderation and monitoring.
- (d) Supports for the development of an equitable system of assessment that is more matched to district and school needs.
- (e) Support for teacher professional development and effective science teaching practices
- (f) Enabling local development and use of high-quality performance tasks for formative and summative classroom assessment

Challenges: Key Features:

- (a) The need to communicate to educators and the public the nature of the new tests and classroom tasks and how they represent an improvement.
- (b) Maintaining comparability and/or equivalence of tasks and scoring across different performance expectations and different tasks, through a moderation and monitoring system
- (c) The costs of supporting local or state development and scoring of periodic classroom performance assessments for grades 3-8.
- (d) The costs of supporting and curating the task and resource bank.
- (e) Developing a matrix sampling approach that provides both valid measurement of district performance and reliable individual student scores.
- (f) The costs of hand scoring of performance tasks.
- (g) Building the competency of science teachers to use assessment at scale.

Appendix A
Stanford NGSS Assessment Project
Assessment Review: Overview
September 2015

The Bechtel team reviewed promising assessment tasks (including selected and constructed response and performance assessments) in order to 1) evaluate the extent to which there are existing tasks that align to the NGSS, at least in part; 2) describe overarching trends and themes in the ways most existing assessments do and do not meet the needs of NGSS; and 3) identify model tasks and describe the aspects of those tasks that are instructive for developing NGSS-aligned assessments .

We began this review by collecting as many promising sample tasks as possible looking both domestically and abroad. The search has been conducted both via internet searches using systematic search terms as well as referral sampling. Our current bank of tasks is not exhaustive nor representative of the entirety of what is out there (we have placed an emphasis on finding innovative items that may be more likely to be aligned with the NGSS). It is intended to provide examples of a variety of assessments--in both form and function--that assessment experts have identified as possible models. We have reviewed a sample of these tasks in light of the NGSS using review categories described below. The findings from this review are described in a separate report (Wertheim et al., in progress).

To date, we have accessed sample tasks from the following sources:

AAAS Project 2061 Science Assessments	Achieve
ActivationLab	Advanced Placement
AQA A-levels	scientificargumentation.stanford.edu
Assessing Scientific Literacy	Biointeractive
Durham University Kind Test	Force Concept Inventory
GISA Science Reading Tasks	Glass Labs games
Innovation Technology in Science Inquiry	International Baccalaureate
Iowa Assessment Handbook	NAEP
Lawson Test of Reasoning	New Zealand Thinking With Evidence Samples
New Zealand National Education Monitoring Project	Oakland Unified End of Course exams
Next Generation Science Assessment Project	ONPAR

Oakland Unified SIRA Items	PISA
PALS and PALM assessments	SCALE
SAVE Science	Test of Scientific Argumentation
Shavelson performance tasks	TIMSS
The Molecular Workbench	University of Cambridge International Exams
U. Wisconsin ConcepTests	West Ed Formative Assessments
University of York EPSE project	
WISE (Berkeley)	

AAAS Project 2061 Science Assessments	Achieve
ActivationLab	Advanced Placement
AQA A-levels	argumentation.stanford.edu
Assessing Scientific Literacy	Biointeractive
Durham University Kind Test	Force Concept Inventory
GISA Science Reading Tasks	Glass Labs games
Innovation Technology in Science Inquiry	International Baccalaureate
Iowa Assessment Handbook	Items from published research literature
Lawson Test of Reasoning	NAEP
New Zealand National Education	New Zealand Thinking With

Monitoring Project	Evidence Samples
Next Generation Science Assessment Project	Oakland Unified End of Course exams
Oakland Unified SIRA Items	ONPAR
PALS and PALM assessments	PISA
SAVE Science	SCALE
Shavelson performance tasks	Test of Scientific Argumentation
The Molecular Workbench	TIMSS
U. Wisconsin ConcepTests	University of Cambridge International Exams
University of York Items	West Ed Formative Assessments
WISE (Berkeley)	

In addition, we have released items from 37 states and from numerous professional journals and research reports.

Categories of Assessment Criteria

We have used the following assessment criteria to aid in our classification and evaluation.

Part 1: Classification

1. **Grade:** What is the grade band for the intended audience?
2. **Item type:** What is the student doing to complete the task?
 - Selected response
 - Constructed response
 - Performance
 - Other
3. **Delivery Platform:** What is the assessment delivery platform?
4. **Timescale:** Approximately how long would it take for students to complete the assessment from start to finish?
5. **Subject area:** If this assessment was developed to assess knowledge or skills for a specific science subject, which subject was it developed for?

Part 2: Evaluation

1. **Engagement:** What aspects of the task allow for choice, self-direction, or collaboration by the students?
2. **Revision:** Do students receive feedback and have the opportunity to revise their responses to the task?
3. **Competencies being assessed:** What competencies or knowledge is *required* to be able to respond to the task correctly?
 - **DCI:** Which Disciplinary Core Idea from NGSS is being probed?
 - **Practice:** Which science practice from NGSS is being probed?
 - **Cross-cutting concepts:** Which cross-cutting concept from NGSS is being probed?
4. **Number of dimensions probed:** How many of the three dimensions of NGSS (disciplinary core ideas, practices, cross-cutting concepts) are being probed?
5. **Integration:** How well integrated are the dimensions represented in the task (none, low, high)?
6. **Common core connections:** Does the task require competencies from Common Core? Check only if Common core connections are explicit in the assessment materials.
7. **Other knowledge or skills required:** Do students have to know or be able to do anything other than competencies that are part of NGSS or Common Core to respond to the task correctly?

8. **Cognitive demand (Using the Depth of Knowledge scale):** How deeply do you have to understand the content to interact successfully with it?
 - Level 1: Recall and Reproduction
 - Level 2: Skills and Concepts
 - Level 3: Strategic Thinking
 - Level 4: Extended Thinking

9. **Validity/reliability info:** Do we have documentation about validity and reliability for the assessment or the instrument that the item came from?

10. **Scoring guide:** Is there a scoring guide, rubric, point system, checklist, a list of criteria for scoring, or a description of levels of students' understanding/skills that accompanies the assessment?

11. **Big ideas:** Rate the degree to which you think the assessment task focuses on the big, essential concepts that are central to the discipline and worth learning and evaluating. (scale of 1-5)

12. **Alignment:** If this task came with information about its alignment to any of the three dimensions, how accurately was its alignment? (none, weak, strong)

13. **Appropriate and accessible:** Provide a rough approximation of how appropriate you think elements of the task are for the intended audiences. (scale of 1-5)
 - Language: Are the words, technical terms, phrasing, and reading load reasonable for all targeted students?
 - Images/diagrams: How appropriate do you think the decoding of graphs and diagrams and interpretation of images is?
 - Equity/cultural sensitivity: Is the assessment sufficiently sensitive to cultural, gender, and other equity concerns?

14. **Student engagement:** Rate the degree to which you think the assessment task is based on activities that are relevant to students. Does it reflect the kinds of questions and activities that would have the potential to engage and motivate students? (scale of 1-5)

15. **Overall evaluation of strengths, weakness, and limitations of the task**

Appendix B:

What are the lessons to be learned from past assessment efforts in California and elsewhere?

B.1. Retrospective survey

SCALE conducted a retrospective research study of a range of performance assessment initiatives that began in the 1990s, tapping into the expertise of those with a deep reservoir of experience from these earlier efforts to integrate the use of performance assessments into large-scale assessment programs in the United States¹¹. The study documents and synthesizes the common political, technical, and practical issues related to these earlier efforts, with the goal of answering three specific questions:

- What were the conditions that helped sustain some of the programs?
- What were the challenges that led to their discontinuation?
- What are some lessons learned that might help inform current assessment initiatives that seek to integrate performance assessment into large-scale student assessment programs?

B.2a. Methods and Data Sources

The research team gathered information about the design, conduct, and outcomes of performance assessment initiatives in the U.S. in the 1990s and in the following twenty years, including both those that have had some longevity as well as those that were quickly dismantled. The study includes a synthesis of research studies and other documentation on those initiatives, as well as the results of interviews with authors and major policy players who championed these innovative state assessment initiatives.

Through an extended literature search of available research papers, solicitation of internal reports from the administrators of the performance assessment systems, and interviews with key stakeholders, we gathered key information about 1) the design of the assessment programs, including evidence about the technical quality of those assessments (reliability and validity); 2) the goals and policy framework of the performance assessments; and 3) the implementation of the assessment programs, including the cost of implementation.

The performance assessment systems that we examined included the following initiatives:

State	Initiative Name	Years of Administration
California	California Learning and Assessment System (CLAS)	1993 – 1994

¹¹ Wei, R.C., Pecheone, R.L., & Wilczak, K.L. (2014). *Performance assessment 2.0: Lessons from large-scale policy and practice*. Stanford Center for Assessment, Learning, and Equity

Connecticut	Connecticut Mastery Test (CMT) Connecticut Academic Performance Test (CAPT)	1985 – present 1994 – present
Kentucky	Kentucky Instructional Results Information Systems (KIRIS)	1991 – 1998
Maryland	Maryland State Performance Assessment System (MSPAP)	1991 – 2002
Nebraska	Nebraska School-based Teacher-led Assessment and Reporting System (STARS)	2001 – 2009
(Multiple states)	New Standards Project (NSP)	1991 – 1999
Rhode Island	Rhode Island Diploma System	2011 – present
Vermont	Vermont Portfolio Assessment Program	1991 – 2004
Wyoming	Wyoming Body of Evidence (BOE)	2001 – present

For more detail, see Appendix C.

B.2b Overview of Findings

In our analysis, we draw parallels between these retrospective lessons gleaned from the 1990s performance assessment initiatives and the conditions today (e.g., policy contexts, technical issues, and practical/implementation issues) to inform our understanding of current challenges and areas for opportunity. The large-scale assessment programs that attempted to include performance assessment emerged because of a growing reliance on assessment as a policy tool to reshape American education and to hold education agencies at all levels accountable for student performance. At the same time, the assessment programs often became the casualties of the same political processes that helped bring them into being when demands for accountability shifted.

Three kinds of lessons learned have emerged from our synthesis of the research that are:

1. Lessons about the role of **political contexts** and the importance of leadership, communication, and public support -.
2. Lessons about **technical quality** and the design of performance assessment systems that support credibility and viability.
3. Lessons about **practical issues** such as cost and implementation factors that supported or hindered the success of performance assessment systems.

B.2c Political Context and Leadership Issues

A crucial factor that either supported or led to the dismantling of large-scale performance assessment programs in the 1990s was the political context in which they were initiated, funded, developed, and implemented.

In our study of the performance assessment initiatives of the 1990s, we identified four major factors related to political context and leadership that shaped the outcomes of the programs:

- I. Shifting purposes for educational assessment
- II. Competing priorities and scarce resources
- III. State politics and educational leadership
- IV. Public acceptance and teacher and parent buy-in

I. Shifting Purposes for Educational Assessment: The Move toward Greater Accountability

Most state assessment systems that used performance assessment methods were developed in the 1990s. Their intention was to use state standards and state testing to drive educational innovation toward more effective teaching. The goals of teaching and learning for those who developed these assessments were broad, but generally incorporated a view that learning goes beyond accumulating knowledge and has the goal of developing students' skills as analytic thinkers and problem solvers. This changed with the passage in 2001 of the act generally known as "No Child Left Behind" which emphasized the need for all students to acquire certain basic skills, intended as a foundation rather than a replacement for the broader skills stressed in prior testing. By making it a federal requirement that every student should receive an individual score in language arts and mathematics in every grade from 3 to 8 and once in high school, and in science once in each school level (K-5, 6-9, and 10-12), the law greatly increased the number of tests to be scored and, given limited budgets, the demand for cheap and efficient testing systems to test the required skills. These tests revealed the unequal outcomes of black and Hispanic students of color and those from low-income homes compared to middle class white students, and tried to force remedies for this problems through goals for improvement and sanctions for schools that did not meet goals. Subsequently, after analysis of this data highlighted the variance of outcomes for different teachers, the stakes of these tests were further raised by federal incentives to include changes in student scores as a part of teacher evaluation measures.

A number of perhaps unintended effects of this regime have been documented. The emphasis on tests of language arts and mathematics has reduced attention and time for to science teaching in elementary and even some middle schools, and the cost and reliability demands of these high stakes tests have resulted in changes away from complex tasks and toward machine scorable (in paper and pencil format) selected response items. These types of item can reliably test a limited set of basic skills and recall of knowledge, but the high stakes emphasis on them has also changed teaching, particularly in low-performing schools, toward more emphasis on the repetitive drill and practice strategies to prepare students for the test, and less emphasis on broader educational goals¹². Part of the reason for new science standards is to recover the teaching of science, and its broader learning goals of building students to think critically and apply their scientific knowledge. Likewise new standards for

¹² Au, Wayne. (2007). High Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, 36(5), 258-267.

Hannaway, J., & Hamilton, L. . (2008). Accountability policies: Implications for school and classroom practices.: Urban Institute: RAND.

mathematics and language arts have led to the development of more varied testing tasks in the new assessment systems aligned to these standards.

II. Competing Priorities and Scarce Resources (Cost of testing)

Another factor that impacted the sustainability of performance-based assessments in state assessment programs in the 1990s was the variability in availability of public and private funding for their development and implementation

In the case of the New Standards Project, philanthropic funds from the Pew Charitable Trusts and the John D. and Catherine T. MacArthur Foundation were used in combination with the project's state membership dues to support development and piloting of New Standards exams in numerous states¹³.

Both philanthropic resources and public funding for education are "soft money"-meaning that the funding fluctuates and, particularly in the case of philanthropic support, almost always disappears after a few years. Both are subject to the booms and busts of the economy, especially at the state level, as well as changing political priorities. When initial seed money is expended, it is often impossible to sustain an expensive education program, especially when the program has insufficient political or public support or experiences any issues of credibility.

During the 2000s, as the demands for student-level accountability increased, the total costs of state testing rose substantially. As a result of No Child Left Behind, state testing costs went from an average of \$8.4 million in 2001 to an average of \$22 million in 2007-2008¹⁴. The federal government funded the increased costs of assessment with an initial 2002 investment of just \$378 million¹⁵ (USDOE, 2013). This meant that states had to reallocate funds from other state education priorities to meet new annual testing demands, and legislators felt pressured to eliminate higher-cost testing programs like those that incorporated performance-based items.

In comparison to state testing programs that exclusively use machine-scored selected-response items, programs that include extended constructed-response items and performance-based items are simply more expensive due to the cost of developing, administering, and hand scoring those types of items. "Typical" assessments (i.e., those with selected-response items only) had an average cost of \$19.93 per student in 2010¹⁶, while "high quality assessments" averaged \$55.67 that same year¹⁷. Studies show

¹³ Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership*, 50(5), 11-15.

¹⁴ Total U.S. spending on standardized tests was almost \$423 million in 2001; for the 2007-2008 school year it was almost \$1.1 billion. Vu, P. (2008, January 17). Do state tests make the grade? *Stateline*. Retrieved from <http://www.pewstates.org/projects/stateline/headlines/do-state-tests-make-the-grade> 85899387452

¹⁵ Section 6113 of the No Child Left Behind Act of 2001 authorized \$490 million to be appropriated for state assessments for fiscal year 2002 (NCLB, 2002), however the final 2002 federal budget included just \$387 million in appropriations for state assessments (USDOE, 2013).

¹⁶ Darling-Hammond and Adamson (2013) argue that the cost of "typical" assessments is actually much higher when the costs of test-prep, benchmark assessments, misdirected classroom instructional time, and other

that, in past initiatives, the cost of scoring performance tasks and on-demand essays ranged from \$1.50 to \$15 per student¹⁸. Faced with increased requirements for testing under NCLB, states made the difficult decision to scale back the proportion of performance-based items in their state assessment programs.

However, it is important to consider not just dollar costs but other costs and benefits when designing an assessment system. Data from machine scored tests is cheaper, but it comes at the price of distorting teaching and learning toward preparation for the limited tasks appearing on the tests, and deeper learning is ignored. To be cost-effective, an assessment system must measure that which is important to learn, not just that which is easiest and cheapest to measure. Cheap but limited data can be very expensive in terms of its impact on longer-term and deeper learning outcomes.

III. State Politics and Educational Leadership

As with any realm of policy, strong support from state leadership at the level of the Governor and Secretary of Education as well as more broadly in the State government are important to the development and funding of any state-level initiative. However, this can also mean that initiatives rise and fall with changes in leadership at the state level, unless they develop a strong and ongoing support constituency that crosses levels of government and political parties. The differences between the states where performance-based testing has survived to the present day and others where it was tried and then abandoned point to the need for a broad-based and well-informed support for the educational goals of this type of testing, as well as careful design to ensure that the testing has the technical characteristics needed for the purposes for which it is to be used.

IV. Public Acceptance and Teacher and Parent Buy-in

Since 2001 policy makers, school administrators, and the public have become used to having a straightforward measure or score which (they believe) can be used to judge the success of their children, or the effectiveness of a teacher, a school, or any educational initiative. The public, and, in particular, the policy makers and the education managers, have been come to value this data, and to trust it). In fact, no single assessment has rich enough information and the validity of standardized tests is questionable. This is particularly true if one wants to ask about a broader range of outcomes than how many students reach a particular level of proficiency in certain basic skills and recalled knowledge.

factors are included in estimates. They argue that the financial cost of implementing systems of performance assessment may actually be *lower* than the financial cost of traditional standardized exams. Moreover, performance assessments support a system of deeper learning while traditional exams divert financial resources towards ineffective teaching and learning practices.

¹⁷ Topol, B., Olson, J., Roeber, E., & Hennon, P. (2013). *Getting to higher-quality assessments: Evaluating costs, benefits, and investment strategies*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

¹⁸ Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Employers who complain that they cannot find job candidates with the kinds of skills they are looking for have led to a renewed emphasis on broader outcomes under the label of “college and career ready” capabilities, and both Common Core and NGSS are driven in part by the need to help students develop these capabilities. The framing of the outcomes in terms of a set of performance standards mean that new standards cannot be fully tested using only selected response tasks. However, public education will be needed before introducing any new assessment system that moves away from machine scored tests of simple but highly limited outcomes. Both the reason for the move to more complex tasks and more complex reports of outcomes will need to be explained. The expectation that scores on the new tests will not match those on prior tests must also be developed and explained. Such an outcome is only acceptable if one believes that the new assessment system will drive the education system towards better outcomes eventually, and will provide fair measures for all students in the system. All segments (policy, education leaders and parents) will need support to reconsider what the data is that they actually need, and what it can and cannot tell them.

B.3 Technical Quality Issues

A second consistent finding from our study of the performance-based assessment initiatives of the 1990s was that the technical quality of the assessments and the way they were scored was generally insufficient for the purposes they were intended to meet. Large-scale assessment programs that use results for accountability purposes and assign high stakes consequences are held to a higher standard of technical quality than those in which the results are used primarily for formative purposes. In a changing policy context, school-level accountability was being significantly intensified, and individual scores for students were expected. However, the performance-based assessment programs that were dismantled near the end of the 1990s and early 2000s had difficulty producing student-level scores that were both defensible and comparable on technical grounds.

There are three main technical quality issues related to performance assessments:

- I. Lack of standardization and comparability across tasks of performance assessments
- II. Validity and content issues
- III. Inter-rater reliability and insufficient item reliability

These technical issues continue to be important considerations in the design of large-scale assessment systems with high-stakes purposes.

I. Lack of Standardization and Comparability of Performance Assessments

A common critique of performance-based and portfolio systems in the 1990s was that it was unclear whether students were being held to the same standard when the content, quality and difficulty of assignments included in the portfolios could not be said to be comparable. On top of that, much of the student work that was entered into the portfolios was revised and polished, the result of peer assistance and teacher feedback. The question of "whose work is it?" became an issue in an era of increasing

accountability in which student scores reflecting unassisted performance were the target of measurement. Student work that was completed with peers or at home with parental assistance became suspect as a trustworthy source of evidence of student learning.

In addition, another reason contributing to the skepticism of many performance assessment systems of the 1990s was that when different performance tasks were administered across students (as was practiced as part of a matrix sampling strategy), it was unclear whether these assessments could be said to be comparable in difficulty. The same critiques about comparability and standardization have been made about the Wyoming Body of Evidence (BOE) system and the Rhode Island Diploma System, which were initiated in the 2000s. Similar criticisms can be found in the UK system designed to measure students' investigative competence¹⁹. This comparability issue also impacts year-to-year or student group to student group comparison of results.

Even when tasks are professionally designed and tested to overcome some of these issues, the very length of the task limits the number of separate topics to be tested for any one student and hence the overall inferences about student capabilities that can be made. However, it is important that the external, high stakes science testing provide good models for student tasks to be used by teachers in both formative and summative assessment as assessment operationalize the construct of what matters. Thus a mixed model which includes a variety of task types and includes a classroom-based performance task as one element appears to us to have the most promise to both provide reliable data and drive classroom practice in the intended direction.

II. Validity and Content Issues

A third criticism of many of the performance assessment initiatives of the 1990s was a perceived focus on process and “soft skills” instead of core content knowledge. The issue here is related to disagreements about what the goals of education should be, the goals of “knowledge” tested by on-demand items testing chiefly factual recall and routine problem solving, or the goal of “career and college readiness” that requires skills and ability to apply knowledge to solve unfamiliar problems. The “performance expectations” of NGSS suggest that what matters is “knowledge in use”, that is not simply declarative knowledge of facts but the ability to engage in science and engineering practices and to use scientific knowledge in a problem-solving or an engineering design situation.

In addition, these tests were criticized as not being valid measures of well-defined student learning. Validity, in this context, is the issue of whether the tests actually measure what they intend to measure. Two issues arise, one that the targeted skills or knowledge is may be poorly defined, and two that success on the task may actually depend more on non-targeted knowledge and skills such as reading or math ability or cultural contextual knowledge. Modern task design helps avoid such problems.

¹⁹ Donnelly, Jim, Buchan, A., Jenkins, E., Laws, P., & Welford, G. (1996). *Investigations by Order: Policy, curriculum and science teachers' work under the Education Reform Act*. Nafferton: Studies in Science Education.

In addition, criticism arose because students showed gains on these measures that were not matched by gains on more traditional tests. However, this can occur when the goals (targeted abilities) of the tests are different. It is quite possible that the gains on one scale – that is on measures the ability to apply knowledge – could occur without gains on the other scale that measures only the quantity of knowledge recalled. As teachers begin to stress the knowledge in use rather than simply declarative knowledge such an outcome is not unexpected. Perhaps because fewer topics are “covered” in such a classroom, scores on a traditional test may decline even though those on a test that requires some of the knowledge to be used in context are improving. What this means is that the issue of validity has multiple aspects, and cannot be settled simply by comparison of one set of results to another when the goals of the two measures are different.

III. Reliability of Individual Student Scores

Another technical quality issue in the 1990s that continues to be a source of critique of performance assessment today is the difficulty of achieving high enough reliability of scores.

If performance assessment scores are used in a high-stakes context (i.e., to hold an individual student or teacher accountable), inter-rater scoring differences become a critical issue. The initial low reliability of local scores that was found in the Vermont and Kentucky portfolio systems is highly related to the training, scoring, and audit systems put in place for the use of hand (human) scoring. Local scoring results can be monitored, and reliability can be improved through regular external audits of local scoring, but it suggests that there are limits to the reliability of local scoring, especially under high-stakes circumstances. For large-scale assessment systems, distributed scoring approaches (blind scoring of randomly assigned student responses) that utilize a cadre of trained scorers from across a state are more likely to lead to greater inter-rater reliability than teachers scoring student responses from their own school or districts. However, such systems add expense.

This kind of distributed scoring system for scoring constructed-response and essay responses has been utilized by large-scale assessment programs and testing companies over the last 15-20 years, producing sufficient levels of reliability for the inclusion of scores in the federally-mandated accountability measures in the new test systems for CCSS. However, the stress must be on the term “sufficient”, experience shows that even with these measures one must expect and accept somewhat lower reliability for these more meaningful tasks. The same will be true for performance tasks in science.

Because performance assessments take longer to administer than traditional forms of assessment, it is unlikely that a large-scale assessment program will include more than one or two performance tasks. The small number of scores generated by a single task, as well as the limited number of performance tasks that can be administered to a single student, limits the reliability and generalizability of those scores. Shavelson, Baxter, and Pine (1991)²⁰ found that a single science performance assessment provides unstable estimates of student performance and recommend that 6-8 performance tasks are

²⁰ Shavelson, Richard J., Baxter, Gail P., & Pine, Jerome. (1991). Performance Assessment in Science. *Applied Measurement in Education*, 4(4), 347 - 362.

needed to improve the reliability and stability of performance assessments as a measure of student learning and ability on a given construct. This puts high demands for testing time!

Some state assessment programs have attempted to address the issue of item reliability by combining performance tasks with selected-response or constructed response items so that the same learning targets, or aspects of those targets, are measured in multiple formats. For example, several states that included performance assessment formats in their assessment programs in the 1990s used a balance of selected-response, constructed-response, and performance-based items within a testing program (e.g., Connecticut's CMT and CAPT; Kentucky's KIRIS; Vermont's Portfolio Assessment Program; the New Standards Project); however, almost none of these systems were able to equate or scale these scores across grade levels (vertical scaling). This must be seen as even more of a challenge for science regardless of test format because there is no well-established sequence of revisiting topics in greater depth, or even of what topics are expected to be learned at what grade level. Thus, with each grade level studying a different mix of topics, the idea of a vertical scaling is difficult to even define.

Finally, a major lesson learned is that it takes time to get these decisions right. Expecting testing programs to have resolved all reliability and validity issues within the first two years of implementation is not a reasonable expectation. This has implications for the timing and phase-in of new assessment programs and its use for high-stakes accountability. While policymakers, school administrators and the public have now become dependent on annual data and thus have a low tolerance for an accountability vacuum, it would be irresponsible for states to use the results of a new large-scale assessment program for high-stakes purposes before the results suggest that such use is technically defensible. Furthermore, it means that year to year comparability of scores will not be re-established until the new system has matured and stabilized. Thus, a timeline for new tests requires not only a pilot stage and a field test stage, but perhaps revisions and a repeated field test before it stabilizes enough to be used for high-stakes purposes. The migration of science testing to computer-based testing opens up one new set of options for the next implementation of science testing, the addition of classroom-based performance tasks to the system could be phased in later, to allow more time to design it well.

As computer technology options develop further, one may wish to introduce further new features (e.g. simulation based tasks where the computer's data on student actions is used to develop elements of the score). There are a number of such approaches currently under development in research settings but perhaps not yet ready for implementation in a large-scale system.

B.4. Practical Issues in Implementing Large-Scale Performance Assessments

A last set of important factors that we found to have an impact on efforts to embed performance assessments into large-scale assessment systems in the 1990s were the practical issues that relate to implementing the assessment systems. This set of factors include and are not limited to: (a) costs and burdens associated with developing, administering, and scoring performance assessments; (b) pressure to quickly scale up and use the assessments for accountability; and (c) the need for a coherent system of curriculum, instructional resources, and professional development.

The higher cost of performance assessment must be set against its value in driving toward instruction and learning that emphasizes “knowledge in use.” Costs will drive the system to a small number of externally-scored performance tasks as an element of a larger system of tasks, combined with matrix sampling methods to monitor whether the full range of standards are being taught. A system in which some elements are used for individual student scores and others for teacher, school or district-level accountability can perhaps allow options that are discarded when the same elements are expected to do both jobs.

As for timeline, in an environment where policy makers require annual data on student performance and consider this data an essential driver of educational reform there is pressure to implement a new system aligned to new standards as quickly as possible. However, past experience suggests this is unwise. It takes some time to develop, test, and debug a new system, and only when that time is invested will the system be able to meet its goals and deliver trustworthy data. The assessments must be seen as part of a *developing* coherent system along with curriculum resources, instructional strategies and professional development for using all of these. One of the lessons of the past is that performance-based assessment systems, or even mixed systems with some performance-based elements, implemented without sufficient preparation, technical quality, connection to other parts of the system, or public education did not long survive. Only in a few states where leadership had the vision to take the time to attend to do all these aspects have such systems continued to this day such as Connecticut (see table above for more examples).

B.5. Recommendations Based on Lessons Learned

All these lessons are salutary; however, they must be set against the lesson that teachers teach to the test so, if only memorized knowledge is tested, then only rote learning will occur. The NGSS standards *are* performance expectations. Assessments that test these expectations will need to overcome the challenges of past performance task assessments. In part this can be done using modern task design methods to ensure sufficient technical validity and reliability, and using the capabilities of computer-delivered tests to allow machine scoring of a richer range of tasks, but the mix of tasks will need to include some performance tasks that cannot be machine scored, although some, such as simulations, may be machine-delivered.

The need to restructure our schools and classrooms to support the acquisition of higher order thinking skills is becoming more urgent every day as the information age is pressuring our educational system to change or be left behind. A principal means to achieve these ends will depend on states and districts moving beyond previous No Child Left Behind (NCLB) policies to rethinking the current structure of the state and national accountability systems that focus primarily on core facts and recall to new systems of assessment that are able to support the development of deeper learning skills that promote broader competencies. Science assessment, in part because it is less emphasized in federally mandated testing, can potentially lead the way by providing a test-bed in which to develop a richer and more coherent assessment approach as one of the drivers to achieve deeper learning for students.

Performance-based assessments require students to use high-level thinking to perform, create, or produce something with transferable real-world application. More than standardized tests of content knowledge, such assessments can provide more useful information about student performance to students, parents, teachers, principals, and policymakers. With the adoption of NGSS there is now a renewed need to develop assessments that include performance assessment components, as well as a richer array of shorter tasks that also probe scientific analysis and reasoning, argumentation and problem solving skills, rather than factual recall. However, an examination of past initiatives suggests that the performance assessment elements of the accountability systems must be carefully used and well-designed. The following are eight key recommendations for successful performance assessment initiatives:

1. Use good task design practices to design assessments that meet intended purposes and meet standards of technical quality, using a mix of short response (selected or constructed), and both simulation-based and classroom-based performance tasks.
2. Design classroom-based performance tasks included as part of the assessment system. These tasks should be curriculum-embedded, and produce well-defined student work-products to be scored following a well-developed scoring rubric. Along with the task and its instructional context, both the required products and the scoring rubric should be communicated to teachers administering the tasks in their classroom. There are three options for scoring such performance tasks. One is train and pay a team to do the scoring, most likely using scorers drawn from known expert teachers. This team then scores student work submitted by the classroom teachers. Second, one can minimize the costs of hand scoring class-room-based performance task products by involving teachers in scoring their own students work. This requires a systematic professional development process before it is instituted, and an ongoing moderation and oversight process. Indeed, such a system has been seen to have significant professional development benefits, leading to improved teaching and learning, where it has been well implemented. Such scores can be sufficiently reliable for use as part of a school and district-level accountability, but not at student-level or teacher level. A third option is to conduct scoring institutes where teachers score student work but not that of their own students. When structured well and facilitated expertly, this can be an excellent professional development exercise and an opportunity for teachers to develop understanding and confidence in the system (especially when such institutes are conducted routinely over time). Professional development credit and/or stipends are typically needed to ensure full participation.
3. Invest in the development of a curated resource bank of high-quality CCSS-aligned performance tasks suitable for use as formative assessment tasks and make it accessible to teachers to support powerful instruction and assessment practices. The resource bank should include rubrics for scoring these tasks, and models for task and scoring rubric development to help teachers in developing further such tasks.
4. Support teachers through professional development to use this resource as part of a coherent system of embedded assessments, curricula, and instructional supports.
5. Minimize the cost of developing performance assessment tasks through economies of scale and cross-state collaboration.

6. Engage with stakeholders more actively, and develop the capacity of educational leaders and policymakers to deeply understand and champion research-based reforms in assessment
7. Engage with the public more actively, and provide timely, accessible information about the new assessment systems and the NGSS.

Appendix C

A Closer Look at Three States' Performance Assessment Programs

In our examination of the nine state level balanced-performance assessment initiatives included in this study, we noted that a few of the initiatives had greater longevity than others. When initiatives did not last more than a few years (e.g., CLAS), this was usually due either to political or leadership changes, or the technical limitations of the assessment (i.e., matrix sampling when student-level results are desired, lack of comparability across assessments) that could not withstand the increased demands for assessment-based accountability. Initiatives that lasted for a longer period of time (more than five years), such as the performance-based assessment programs in Kentucky, Maryland, Connecticut, and Wyoming, experienced success due to the continuity of political leadership within the state, the overall technical quality of the assessment, and the level of buy-in from teacher and other stakeholder groups.

One state, in particular, Connecticut, stands out in terms of the longevity of its assessment system. While the Connecticut Mastery Tests and Connecticut Academic Performance Test have evolved over the last 25 years – with some of the on-demand classroom-based performance items being eliminated – the state has been able to sustain a high quality assessment that continues to incorporate performance-based items along with selected-response and short constructed-response items.

California Learning Assessment System (CLAS)	
Duration	1993–1994
Grades Tested	4, 8, 10
Content Areas	Reading, writing, mathematics, science, social studies
Description of Assessment	Multiple choice, constructed response, and performance tasks
Technical Characteristics	Matrix sampling
Scoring	Scored in-state by teachers
Score Reporting Level	School-level reporting
Accountability System/Purpose of Assessment	No formal consequences were attached to CLAS results; school-level rewards and sanctions were not tied to assessment performance. CLAS was initially designed to give curricular information to schools, districts, and the state.

State Standards/ Frameworks	California Frameworks
Current Status	CLAS was discontinued in October 1994 due to its inability to produce individual-level score reports and mounting opposition to the cost and content of the exam.

Connecticut

Connecticut Mastery Test (CMT) & Connecticut Academic Performance Test (CAPT)	
Duration	CMT: 1985 – present CAPT: 1994 – present
Grades Tested	CMT: 3–8 CAPT: 10
Content Areas	CMT: Mathematics, reading, writing, science (science added in 2008) CAPT: Mathematics, reading (interdisciplinary), writing (interdisciplinary), science
Description of Assessment	CMT: Selected-response and open-ended items, essay responses CAPT: Selected-response and open-ended items, essay responses, questions related to curriculum-embedded performance tasks, on-demand performance tasks (eliminated in 2007)
Technical Characteristics	Criterion-referenced. Scale scores within a grade and content area comparable from one year to the next; scale scores not comparable across grade levels.
Timeline	Administered in March, scores released in August
Scoring	Scored by Measurement Incorporated; prior to 1992, CMT scoring was conducted within the state
Score Reporting Level	Individual student score reports; school and district summary results
Accountability System/Purpose of Assessment	CAPT and CMT were designed to be low stakes assessments. Stakes have risen as both tests are now used to meet federally mandated requirements. CAPT scores are included in district/school graduation criteria, but cannot be the sole criteria for graduation.
State Standards/ Frameworks	<i>The Connecticut Framework: K-12 Curricular Goals and Standards</i>
Current Status	CMT and CAPT will be administered for the final time during the 2013-2014 school year; Connecticut will begin using assessments from the Smarter Balanced Assessment Consortium (SBAC) in 2014.

Kentucky

Kentucky Instructional Results Information System (KIRIS)	
Duration	1991 – 1998
Grades Tested	4, 5, 7, 8, 11 ²¹
Content Areas	Reading, writing, mathematics, social studies, science, arts and humanities, practical living/vocational studies
Description of Assessment	Selected-response items, open-ended written tasks, performance events, and a portfolio (math, writing) reflecting a student's best work ²²
Technical Characteristics	On-demand test components administered through matrix sampling; portfolios scored holistically
Timeline	Assessment administered in spring with results reported annually; schools formally evaluated every two years (accountability cycle)
Scoring	Portfolios in writing and math locally scored by teachers with a sample sent to the state for rescoring to establish reliability; on-demand components scored by outside testing company
Score Reporting Level	School performance data; individual student scores not released
Accountability System/Purpose of Assessment	KERA (Kentucky Education Reform Act) instituted a school accountability index comprised of KIRIS results and non-cognitive measures (dropout rates, attendance rates, etc.). KIRIS results accounted for five-sixths of each school's score. Each school received an overall score on a scale of 0-140; all schools were expected to meet the long-term goal of at least 100 at the end of 20 years. Schools that reached or exceeded their short-term target score could receive monetary rewards; sanctions for schools that failed to reach their target score included state takeover, mandatory School Transformation Plans, or intervention by a "distinguished educator."
State Standards/Frameworks	<i>Transformations: Kentucky's Curriculum Framework, Volume I (1993) and Volume II (1995), Core Content for Assessment (1996)</i>
Current Status	Due to high costs and mounting political opposition, KIRIS was dismantled in 1998 and replaced by the Commonwealth Accountability Testing System, which included writing portfolios and on-demand testing components, including a shorter writing task and selected-response items.

²¹ KIRIS originally tested students in grades 4, 8, and 12 in reading, writing, social science, science, mathematics, arts and humanities, and practical living/vocational studies. Assessments were divided between grades 4/5 and 7/8 beginning with the 1996-97 school year. In grades 4 and 7, students completed on-demand assessments in reading, science, and writing, plus a yearlong writing portfolio; in grades 5 and 8, students completed on-demand assessments in math, social studies, arts and humanities, and practical living/vocational studies, plus a yearlong portfolio in math. Testing was moved from 12th to 11th grade in 1995 (Koretz & Barron, 1998; NRC, 2010; Stecher, 1997).

²² Test composition changed several times, and not all task types were included each year. The on-demand open-ended writing task was added in 1997. Multiple choice items were eliminated in 1995 but reintroduced in 1997. On-demand performance tasks across all five subject areas were dropped in 1996.

Kentucky was one of the first states to implement a comprehensive state education accountability system. The Kentucky Instructional Results and Information System (KIRIS), the state’s assessment system, included a balance of item formats, including selected-response questions, on-demand and curriculum-embedded performance tasks, and a portfolio component.

Maryland

Maryland State Performance Assessment Program (MSPAP)	
Duration	1991 – 2002
Grades Tested	3, 5, 8
Content Areas	Reading, writing, language usage, mathematics, science, social studies
Description of Assessment	8-10 on-demand performance tasks (some interdisciplinary), including pre-assessment group activities with manipulatives. No selected-response items.
Technical Characteristics	At each grade level, 20 tasks were used to assess school performance across the six content areas. Students were assigned on a random basis to one of three test form “clusters.” Each cluster included just 8-10 of the grade’s tasks, meaning that each student did not complete all 20 of their grade’s tasks.
Timeline	Assessment administered in May; score reports released in November
Scoring	Scored by Maryland teachers; teacher scoring procedures were moderated through check sets, accuracy sets, and spot checks
Score Reporting Level	School performance data; individual student scores not released
Accountability System/Purpose of Assessment	Schools were expected to meet standards for satisfactory performance by 1996 (later changed to 2000). A school was rated satisfactory if 70% or more students scored level 1, 2, or 3 on MSPAP’s five-point scale.
State Standards/Frameworks	Maryland Learning Outcomes
Current Status	MSPAP could not feasibly meet the requirements of NCLB; it was replaced by a more traditional on-demand assessment in 2002.

The Maryland State Performance Assessment Program (MSPAP) was an entirely performance-based assessment consisting of interdisciplinary performance activities and extended, multi-part tasks. The assessment met standards for reliability and validity at the school level but was

ultimately discontinued because it could not technically and financially provide the individual student score reports required by NCLB.