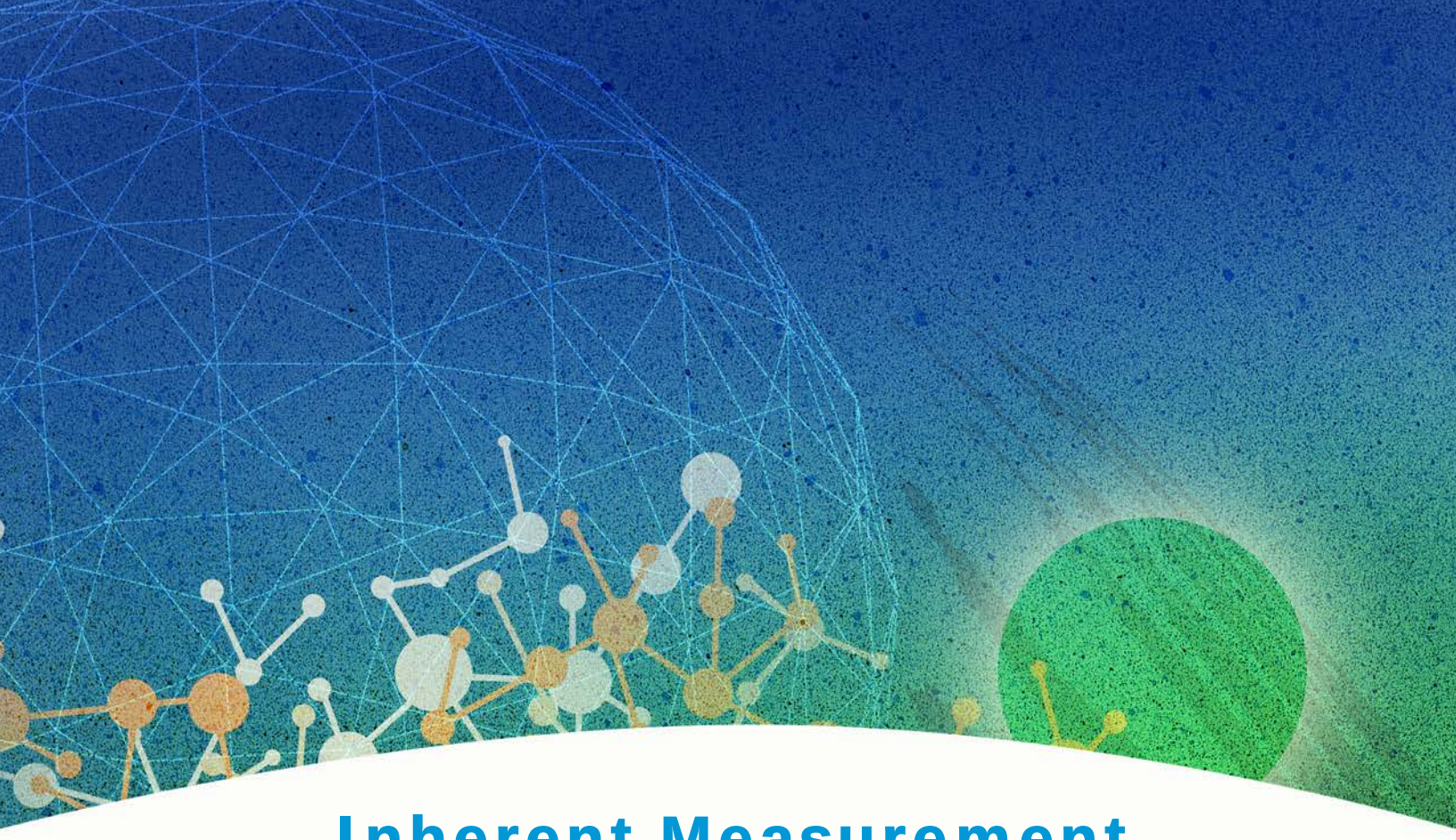

Inherent Measurement Challenges in the Next Generation Science Standards for Both Formative and Summative Assessment

Joanna S. Gorin and Robert J. Mislevy

September 2013

K-12 Center at ETS

Invitational Research Symposium on Science Assessment

K-12 Center at ETS
Invitational Research Symposium on Science Assessment

Inherent Measurement Challenges in the Next Generation Science Standards for Both Formative and Summative Assessment

Joanna S. Gorin and Robert J. Mislevy
Educational Testing Service, Princeton, New Jersey

The Next Generation Science Standards (NGSS) are ambitious in their goals for and demands of teaching and assessment, to be sure. Existing science standards are generally believed to lead to curriculum and assessments that have failed in preparing high school graduates to appreciate science; engage in public discussion of scientific issues; carefully consume and relate scientific knowledge to their lives; and, most importantly, acquire the necessary skills (or desire) to enter and succeed in science, technology, and engineering careers (Carnegie Corporation of New York & Institute for Advanced Study, 2007; National Research Council [NRC], 2007; Schmidt, Houang, & Cogan, 2002). As compared to these earlier standards, NGSS deemphasizes discrete facts taught and assessed in de-contextualized experiences that stifle engagement and limit connections to students' real-world problems and activities. Further, the NGSS consider science education from a developmental perspective, one that sequences curricula over multiple years, rather than as separate standards for each grade that are often disjointed and incoherent in their broader goals. The conceptual shifts implied by these overarching changes in the NGSS have been summarized as follows:

1. K–12 science education should reflect the real-world interconnections in science.
2. Use all science practices and crosscutting concepts to teach all core ideas all year.
3. Science concepts build coherently across K–12.
4. The NGSS focus on deeper understanding and application of content.
5. Science and engineering are integrated in science education from K–12.
6. Science standards coordinate with English Language Arts and Mathematics Common Core state standards.

Taken individually or as a whole, one is struck by the complexity of the NGSS—both the complexity of the standards themselves and the complexity of their implications for assessment, particularly summative assessment for individual high-stakes accountability reporting. The motivation for the NGSS was to design the framework and science standards for instruction and assessment to reflect the inherent complexity in scientific understanding and reasoning as it exists in the real world. The

increased complexity of the NGSS was purposefully designed to get away from the relative simplicity of the previous standards (e.g., discrete, decontextualized knowledge; year-by-year, unconnected standards). The earlier, less-complex standards led to overly simplified instruction and assessment that did not support the desired outcomes, namely students with the ability to reason, function, and perform in scientific contexts. While the purposeful complexity of the NGSS has intended positive effects on the validity and utility of scores for improving science learning, it inevitably introduces significant challenges for some forms of NGSS assessment. These standards reflect the complexities of modern science, specifically through the new call for science instruction and assessment to always intertwine the three NGSS dimensions. This is a new demand for assessment and will be difficult to accomplish in summative assessments with individual reporting given the prevailing/reasonable state parameters for acceptable testing time and cost. In this paper, we discuss three types of challenges for NGSS assessment resulting from the complexity in the framework and standards themselves—challenges for task design and scoring, challenges for psychometric modeling, and practical and logistical challenges. We then propose three general strategies for successful NGSS assessment design: (a) build a coherent system of assessments, (b) borrow information from available sources to support and simplify the assessment design, and (c) clearly articulate and evaluate assessment design choices relative to the assessment goals.

NGSS Assessment Use Cases and Challenges

As assessment developers, our challenge is to develop an assessment that meets as many of its goals as possible, in the simplest way possible, which may not be very simple at all. Consider the challenge posed to the National Assessment of Educational Progress (NAEP) in 1983 when a policy change called for scores to be reported at the latent trait level, rather than at the item level. The multistage scaling solution based on complex sampling, spiraled booklet administration, complex conditioning models, and plausible values, which continues to shape NAEP designs today, were far more complex than what was implemented with other large-scale testing programs at the time (Beaton & Zwick, 1992; Johnson & Rust, 1992; Mislevy, Johnson, & Muraki, 1992; Rust & Johnson, 1992; Yamamoto & Mazzeo, 1992). In fact, the assessment methodology was so complex that two independent studies were commissioned to examine whether such a complex approach was needed. The conclusion of these studies was that, given what the policy makers wanted to report on, how they wanted to use the scores, what could be supported financially in terms of sample size, and other factors—yes, the complexity was, in fact, necessary. If one wanted a simpler assessment system and methodology, then some of the goals, objectives, or constraints on the system would need to be sacrificed. The decision is thus one of priorities and values. If you value all of the system goals equally, then the result may be a very complex solution; if you value a simple solution (i.e., a simple system), then the goals themselves might need to be simplified.

Educational assessment increasingly must serve multiple stakeholders, each of whom wants to use assessment results for different purposes. Stakeholders include students, teachers, administrators,

policy makers, politicians, and researchers, any of whom may want to use assessment scores in different ways. For example, as part of their responsibilities for school improvement, policy makers rely on assessment results for decisions on various issues ranging from professional development and staff allocation, to program and curriculum selection and evaluation, to monitoring and improving student achievement. Herman (2010) lays out the wide array of educational test options that emerge when considering the type and form of assessment (annual, through-course, school/district, classroom), the type of assessment (end of year, end of course/unit/semester, benchmark, and formative), primary user (state, district, school, teacher, parent, student, public), and finally specific use (e.g., teacher/principal/ school effectiveness, assigning grades, identify struggling/promising students, inform short-term learning). Each combination of stakeholder and purpose defines the specific assessment needs and constraints, which suggest particular forms and types of assessment that are more or less appropriate.

We propose that all of the challenges for NGSS complex assessment are consequences of both the increased complexity of the NGSS standards themselves and the need to serve increasingly varied uses and stakeholders. A significantly problematic challenge in meeting the needs and constraints for an assessment use or a stakeholder may be of little significance for another. Therefore, we begin by making more explicit the varied uses, purposes, and contexts that NGSS assessment must serve as a basis for discussion of the challenges themselves.

NGSS Assessment Use Cases

A *use case* in systems design describes the actors, information, and processes involved in meeting some recurring function, like withdrawing cash from an ATM or updating a customer database. An *assessment use case* describes a configuration of actors, information, and processes that serve a recurring assessments purpose. Table 1 lists five use cases that can be envisioned for assessments based on the NGSS, each of which differs on one or more dimensions from one another to varying degrees. We will use them to bring out some measurement challenges that arise in the form of design trade-offs, since in them different people need different kinds of information, have different background information, and have different objectives. While each of the use cases differs on one or more dimensions, stark differences between Use Case 1, “large-scale ‘drop in from the sky’ accountability tests,” and Use Case 4, “formative assessment in a classroom,” make these two particularly useful to illustrate why no single assessment will do a good job for all the purposes people might have in mind for NGSS. Drawing primarily from these two most common and most dissimilar use cases, we will explain how assessments designed to optimize their value for different use cases can look very different from one another, yet all still be consistent with NGSS’s view of the nature of capabilities in science and how students develop it. We refer less to the other three use cases, as in some senses they represent special cases of the formative or summative use cases for which the challenges are mitigated or exacerbated by the unique characteristics of that assessment context or purpose.

Table 1. Five Use Cases for NGSS Assessment

Use case	Key features with measurement implications
Large-scale accountability tests (States' RTTT tests)	***Are not* directly connected to what students are studying** (*drop in from the sky*, or DIFTS, from students' perspective). Can be high stakes for students, teachers, and/or schools. Address standards targeted to grades. Basically (samples of) the same material in the same way to all students at the same time. Administered at a time chosen by the educational agency (e.g., state), usually toward the end of the school year. Student-level scores (?) Comparable across students and schools. Limited usefulness to individual students' learning.
Large-scale educational surveys (e.g., the National Assessment of Educational Progression, or NAEP)	***Are not* directly connected to what students are studying** (DIFTS). Low stakes for students, teachers, and schools. Address content framework that overlaps but need not be the same as standards targeted to grades. Basically (samples of) the same material in the same way to all students at the same time. Administered at a time chosen by the (external) program. No student-level scores. Comparable across states and reporting groups. Not useful to individual students' learning.
Summative assessment to assign grades/course credit (i.e., end-of-course tests)	***Are* directly connected to what students are studying.** High stakes for students. Can address NGSS content standards and performance expectations, at grade level or not, as appropriate to students and course. Administered at end of a course that students have studied. Results not generally comparable across students and schools in the sense of who and when, even if common forms. Useful to evaluate what individual students have been studying. Useful target for student learning but otherwise not useful to aid learning along the way.
Formative assessment in classrooms (e.g., quizzes, feedback during a project, continuous evaluation in an online learning system)	***Are* directly connected to what students are studying.** Low stakes for students. Can address NGSS content standards and performance expectations, at grade level or not, as appropriate to students and course. Administered when, how, and to whom it is most useful to guide learning.

Use case	Key features with measurement implications
	Results not generally comparable across students and schools in the sense of who and when, even if common forms. Aiding individual students' learning along the way is *raison d'être*.
Research to evaluate interventions, curriculum, and policy	**May or may not be directly connected to what students are studying—depends on research objective.** Usually low stakes for students, teachers, and schools. Can address NGSS content standards and performance expectations, at grade level or not, as appropriate to research objective. Results usually generally comparable only within research study. Only aids individual students' learning along the way if it is pertinent to the research study.

Note. NGSS = Next Generation Science Standards; RTTT = Race to the Top; DIFTS = drop in from the sky; NAEP = National Assessment of Educational Progress.

Use Case 1: End-of-year summative accountability assessments for administrator reports. Annual end-of-year summative assessments for accountability are universal in K-12 education for high-stakes decisions about school effectiveness. Standardized tests with common forms are administered universally to all students across a given state on the same day, at the same time, with the assumption that students have had equal opportunity to learn the standards covered by the tests. Mislevy has referred to such context-free assessment opportunities as *drop in from the sky* (DIFTS) assessments (Braun & Mislevy, 2005). DIFTS are characterized by lack of contextualization with respect to both individuals' learning histories and their learning environments. DIFTS neither consider whether a student has received appropriate instruction on prerequisite skills, nor that student's motivation and engagement in learning and completing the assessment. This is essentially the assessment context where we know nothing about students except for how they behave on and respond to an assessment. Specifically, scores from accountability tests are intended to be representative of knowledge and learning across the entire curriculum for a given discipline in a given year. The validity of these scores for the intended accountability purpose—to identify students, classrooms, and schools where teaching is not leading to learning—is predicated on an assumption that every student in the state (or in a group of states) has been studying the same content in a given grade. This flies in the face of what most educators know, which is that not even every child in a single class is maximally benefiting when everyone in the class is studying the same thing, at the same time and pace. Thus, very effective teachers might design classroom instruction in a way that invalidates the inferences from our DIFTS accountability tests. Still, scores from these tests are generally used to document the status and/or growth of students' knowledge absent any information about classroom experiences or opportunity to learn. With the high stakes associated with scores in this use case, the design requirements have typically emphasized reliability and construct representativeness—with specific focus on curriculum and content representation of the tests relative to the grade-level standards.

Use Case 4: Formative assessment for teachers to plan classroom instruction. Formative assessment is a use case of growing interest to K-12 educators and policy makers (Heritage, 2010). Formative assessment use cases call for a variety of assessment tools, including tests, observations, and other data sources, which are used to identify where a student is in learning relative to where the student wants to be. Further, formative assessment results are expected to provide information about how to move the student forward in learning through instruction or other learning activities. While reliability and validity are important here, as with accountability assessment, the diagnostic nature of formative assessment use requires greater attention to the sensitivity of the test items to fine-grained models of learning at critical points in skill and knowledge development (Heritage, 2008). Tests designed for formative uses are, thus, more likely to require items that probe more deeply into focused sets of skills, emphasizing depth versus breath of content, than tests designed for accountability uses. In this use case, much more is known about both the student's past learning and what instructional options are available for what to do next. Neither is apparent when we look at the assessment per se—what is on

the paper or the computer screen. We will see, however, that these factors have great impact on the evidentiary value of an assessment for a given purpose in a given context.

Having laid out the key NGSS assessment use cases, we turn to discussion of the specific challenges introduced for assessment developers and consumers. The discussion is structured around three general categories of challenges—task design challenges, psychometric challenges, and practical challenges—summarized in Table 2. These challenges manifest differently and take on unique significance in the different use cases, making generalized discussion of their implications and solutions difficult. Each use case demands particular assessments as artifacts to be designed for particular purposes in various feedback loops, and each use case presents its own profile of purposes, constraints, resources, and contextualization. Thus, a considerably problematic challenge in one use case may be of little significance for another due to the defining attributes and design features that characterize the use case itself. Further, while our discussion is structured to address each individual challenge separately, the reader will quickly find that they are highly interrelated. Solutions to overcome one challenge will immediately have implications for or be affected by another; thus, we will necessarily discuss the relationships among the challenges.

Table 2 Primary Challenges for NGSS Assessment

Task design and scoring challenges	Psychometric challenges	Logistical and practical challenges
1. Appropriate construct models.	1. Dimensionality assessment with complex structure.	1. Limitations on testing time.
2. Complex assessment tasks.	2. Scaling and estimation.	2. Increased test development costs.
3. Critical role of technology in design and scoring.	3. Reliability and generalizability requirements.	3. Technology requirements for administration.
4. Larger amounts and more varied types of data.	4. Multidimensional vertical scaling and construct shift.	4. Accessibility and universal design.

Task Design and Scoring Challenges of NGSS

Choices about item and response format should ideally be driven by the nature of the behavioral evidence we need to support our assessment claims. That is, if I want to make claims about students' reasoning about the use of models, then I should design a task that provides evidence about how a student uses a model to solve a science problem. One of the most common criticisms of traditional educational tests is their predominant use of multiple-choice (MC) and other forced-choice response formats for measuring targeted constructs. Considerable debate surrounds the question of whether these traditional assessment tasks provide evidence to support claims about higher order

thinking skills (Gorin & Svetina, 2011; Haladyna, 2004). While these items *can* be designed to reflect higher order reasoning and cognitive processes[1], more often than not, MC items, as one usually encounters them, tend to elicit student responses that are evidence of discrete knowledge, rote memorization or recall, or low-level cognitive reasoning.

So why, one might ask, are MC questions the dominant item format on educational tests? First and foremost, because the assessment community understands how to write them, score them, and psychometrically model them. History has shown that well-written MC items can yield internally consistent, unidimensional scales that encompass a wide range of content in a relatively short testing time. Fundamental assessment development principles require multiple bits of evidence about a construct to achieve sufficient score stability and generalizability to support associated high-stakes decisions (Green, 1978). Thus, MC items fit one traditional educational assessment design challenge—to achieve high reliability for individual student scores representative of a large number of discrete content and performance standards.

As education standards have increasingly emphasized higher order cognitive processes, assessment developers have responded by including short and extended constructed response items, most notably essays, on their high-stakes assessments. Both in terms of face validity and construct validity, the use of essays and short-answer items is appealing in that the assessment tasks and associated scoring rules are more obviously aligned with the targeted reasoning and higher cognitive skills. However, even with the more extended response formats offered by these items, several inherited attributes of traditional assessment tasks persist. First, most high-stakes tests are administered via paper and pencil. This restricts the type of stimuli that can be presented in item stems, as well as the types of behaviors that students can demonstrate in their responses (Quellmalz & Haertel, 2004). Second, performance items, while written to capture more cognitively complex constructs, are still designed to optimize other traditional design features. For psychometric reasons, items are written to be statistically independent. In practice, this results in multiple relatively decontextualized (or at least contextually distinct) items, each of which is treated as a separate problem-solving task. Given that the majority of real-world problem-solving activities are highly contextualized, dependent upon one another, and multidimensional in nature, the artificial tasks that make their way on to traditional educational tests are unrecognizable in terms of their real-world significance. To the extent that the constructs measured by these items are a function of task design, we miss the mark of measuring the real-world knowledge, skills, and abilities that we sought to assess with the use of performance tasks (Champagne, Kouba, & Hurley, 2000; Quellmalz & Haertel, 2004).

[1] For example, in Microsoft and Cisco certification exams, there are tasks that require relatively short explorations and interactions in a simulated network environment, and the examinee is asked to answer a set of MC items about properties, problems, solutions, interaction characteristics, and so forth of the network. Such MC items capture evidence of high-level reasoning and strategies, and involve interaction and cycles of inquiry.

The initial challenge for NGSS measurement is, therefore, to design tasks that elicit the rich cognitive processes that define the hard-to-measure constructs as they were conceived and drafted by the standards' authors. This requires, first, that we have sufficient understanding of the construct itself, at an appropriate grain size, to design tasks that engage students in the appropriate knowledge, skills, and abilities (KSAs). Then we must have the capability to design tasks that elicit the appropriate set of skills. With respect to the appropriate model of the construct, the NGSS themselves provide descriptions of the intended KSAs at a particular grain size. However, it is unclear whether they are sufficiently detailed or presented in a manner that can be easily translated into assessment tasks. It is more likely that, to generate appropriate assessment tasks of the intended NGSS KSAs, more cognitively rich models of the constructs are needed (Gorin, 2006; Gorin & Embretson, 2012; Leighton & Gierl, 2011).

Competency models, learning progressions, and other emerging knowledge representations from the cognitive and learning sciences offer significant promise for complex task design (Briggs, Alonzo, Schwab, & Wilson, 2006; Corcoran, Mosher, & Rogat, 2009; Gorin, 2006; Heritage, 2007, 2008). Whereas performance standards as a basis for item and task development are typically framed in terms of behavioral outcomes, learning progression and the like describe the nature of cognitive processes, the representations that students hold about a problem or discipline, and the habits-of-minds that govern how students approach the problem-solving endeavor. These more cognitively based models are most critical, particularly for the formative assessment use case in which the goal of the assessment is to use student performance data to make fine-grained inferences about their current learning and to plan future learning activities. Given the historical emphasis on summative assessments for accountability, it is not surprising that the behavioral grain-size inherent in traditional construct definitions and performance standards have been sufficient for assessment task design. Perhaps for the accountability use case, the more traditional task types aligned to behavioral performance standards are sufficient. This would then imply that we need multiple representations of our constructs, at multiple grain sizes, to support multiple task types, each suited to its particular purpose—a view consistent with the notion of a system of assessments as opposed to a single test to serve all purposes.

The second task design challenge is that the tasks must be designed to give rise to observable pieces of evidence (i.e., data) that can be scored in terms of the complex claims (i.e., multidimensional, hard-to-measure constructs). Response formats for various item types have been compared by several researchers in terms of their level of constraint imposed on the students' response (Scalise & Gifford, 2006; Wilson, 2004). The most constrained items, those with fully selected response formats, like MC and true/false items, are easily scored but provide a single, limited data point from which we can make inferences about students' knowledge. The least constrained tasks, those with "fully constructed" response types, include projects, portfolios, interviews, and performances (Scalise & Gifford, 2006, p. 5). The responses captured in the less-constrained formats provide more varied evidence sources, including those associated with the more cognitively complex KSAs targeted in the new standards. That said, the more open response formats pose challenges of their own in terms of scoring and reporting, a fact that

has motivated an entire field of research on human and automated scoring of essays on high-stakes tests.

While it is unlikely that there is a single solution to the NGSS assessment task design requirements, research and development in the learning sciences would suggest one certainty—the critical role of technology. Moving away from the traditional paper-and-pencil format into a digital assessment platform releases many of the constraints highlighted above. Through technology-enhanced assessment tasks, rich stimulus materials including multimedia, virtual reality, and multimodal capabilities can be provided to set the problem-solving stage for an assessment activity. Research in the cognitive and learning sciences has produced innovations in technology-enhanced learning tools to address the shortcomings of traditional standards-based instructional design. Digital environments have been designed to create authentic, engaging, and challenging learning activities. These interactive learning tools, including simulation systems, intelligent tutoring systems, and educational games can provide the multidimensional, contextualized problem space that adheres to the rich set of KSAs like those of the NGSS (Behrens, Mislevy, DiCerbo, & Levy, 2012; Rupp, diCerbo et al., 2012; Rupp, Gushta, Mislevy, & Shaffer, 2010). Further, by virtue of their digital design, they can capture more, and more varied, data during and at the end of student problem solving (Rupp, Nugent, & Nelson, 2012). The variety of innovative response formats increases significantly just with the availability of a few basic technologies such as drag and drop, highlighting, equation editing, and graphing. At its most extreme, the notion of a *response format* is rendered irrelevant in that data from student interaction is streaming in real time with every click, keystroke, or other interaction being recorded by the technology. DiCerbo and Behrens (2011) write of the "digital ocean" that is available from technology-enhanced learning and assessment activities that can be mined as evidence of our targeted KSAs. There are seemingly limitless data that could be mined to inform and predict student learning, all seamlessly embedded within very naturalistic assessment and learning activities (Shute, 2011).

Not only can data from these and other more open-ended tasks be captured automatically, but also there is increased potential for automated scoring of responses (Williamson, Mislevy, & Bejar, 2006). Returning to Scalise and Gifford's (2006) taxonomy of response formats, they specifically point out cells for which automated scoring is feasible. As automated scoring technologies continue to develop and improve, even more cells in this matrix will become automated scorable. Designing assessments that maximize use of auto-scorable formats could get the most bang for the buck in evidentiary value, for a given cost. Significant effort has gone into automated scoring of essay-like responses because they are both important and familiar, but other applications in non-essay formats can be even more advantageous (Williamson et al., 2006). This is especially so in science, since so much of scientific reasoning is carried out through representations and symbol systems. Student diagramming, concept mapping, modeling, data collection, and even simulated laboratory experimentation all have the potential for automated scoring, which could serve to offset the cost of the increased complexity associated with complex technology-enhanced tasks.

With all of the research and development efforts on technology-based tools for learning hard-to-measure constructs in science, NGSS complex task design seems a surmountable challenge. The tools and technologies emerging from the cognitive and learning sciences to improve instruction and student learning should lend themselves quite nicely to the formative assessment use case (Heritage, 2008, 2010; Quellmalz & Haertel, 2004; Scalise, 2012). In formative assessment, the intent is for scores to be used as feedback to teachers and students to direct subsequent learning activities. The types of rich learning activities that are likely to provide evidence for this purpose are, by design, highly multidimensional and deeply contextualized within the classroom context and the embedded curricular program. The assessment tasks must be designed to elicit evidence aligned with models of student learning and teachers' instructional practices. Furthermore, the evidence from these tasks must be effectively communicated to teachers and students in a way that can support decisions about how to proceed in learning and instruction. With the lines between formative assessment, learning, and instruction blurred, these tools are obvious choices for NGSS assessment tasks that provide instructionally relevant information for students and teachers about the hard-to-measure constructs of interest.

The accountability use case is another matter. In the accountability use case, the psychometric and validity requirements of content coverage, correctly specified dimensionality, scaling, equating, reliability, and generalizability are of relatively larger importance than the issue of instructional relevance, as was the case in the formative use case. Whether and how the proposed complex, context-rich, technology-enhanced tasks can be used in the DIFTS context of accountability assessment is the largely unanswered question. The challenge in this use case stems in part from the ambitious breadth of content and skills most end-of-year assessments target, as well as the fact that we have little context or information about students other than the data generated from the assessment itself. In the formative use case, scores can be interpreted with full knowledge of the specific content, context, and format of instruction. For summative accountability testing, we have no such luxury. Rather, either we must assume that instructional content, context, and format (including the use of technology-supported tools) is sufficiently equivalent across students to have negligible effects on performance, or we must design assessments that sample broadly to average out any differential effects on individual student test performance.

NGSS Performance Expectations

The effect of content, context, and format on the generalizability of score interpretations for particular assessment tasks can be tempered or exacerbated depending on the degree to which these factors vary across and within instructional and assessment activities. The NGSS architecture intentionally gives considerable latitude for instructional and assessment design choices through the use of *performance expectations* as the definitional form of the standards. Performance expectations are the assessable statements of what students should know and be able to do and are written to combine

the three NGSS dimensions. While they provide descriptions of the achievements all students should be able to demonstrate, they do not translate directly into any single instructional activity or assessment task.

In drafting the standards, NGSS authors initially sought to define performance expectations combining the practices and cross-cutting concepts within each appropriate disciplinary core idea as explicitly stated in the framework. However, using verbatim the detailed developmental progressions used to define the core ideas resulted in "bulky" statements that were difficult for readers to interpret and understand how to apply. Ultimately, the NGSS authors elected to write performance expectations "to communicate a 'big idea' that combined content from the three foundation boxes" (NGSS Lead States., 2013b, p. 2). For defining learning outcomes and goals as intended by the performance expectations, this is understandable– we want to know that students emerge from science and engineering education with competency in the key practices and concepts as they interact with core disciplinary ideas. However, while performance expectations as definitions of learning objectives may focus on "big ideas," instructional and assessment activities designed to mimic real-world problem solving require specificity with respect to context and format. We can encounter serious limits when we define skills in a decontextualized way and assess them as such or make inferences from performance in one domain to a different domain. This is manifest as *low generalizability* in the measurement frame, a point we return to shortly. To assist in making decisions about specific instructional and assessment tasks, the NGSS includes c*larification statements* for many of the performance expectations provide some guidance as to some of the contexts in which one might develop activities measuring the expectations. These statements serve to highlight the fact that there are a variety of contexts, each with its own context-specific content knowledge, in which one might choose to teach or assess the same expectation across classrooms. Using these clarification statements in conjunction with the performance expectations, and the individual descriptions of the targeted practices, ideas, and concepts, choices will need to be made about the particular instructional or assessment tasks.

Let us consider a specific performance expectation within the Earth and Human Activity Disciplinary Core Idea (DCI) for the 3rd–5th grade NGSS band: 4-ESS3-1. Obtain and combine information to describe that energy and fuels are derived from natural resources and their uses affect the environment. In addition to three Earth and Human Activity DCIs—Natural Resources, Natural Hazards, and Designing Solutions to Engineering Problems—this performance expectation was developed to incorporate two practices—(a) Constructing Explanations and Designing Solutions and (b) Obtaining, Evaluating, and Communicating Information—as well as two crosscutting concepts—(a) Cause and Effect and Interdependence of Science, Engineering, and Technology and (b) Influence of Engineering, Technology, and Science on Society and the Natural World. To measure this or any performance standard, a variety of assessment tasks could be design, each of which differs in terms of the task format and scientific problem solving context. The clarification statement for performance expectation 4-ESS3-1 reads

> Examples of renewable energy resources could include wind energy, water behind dams, and sunlight; non-renewable energy resources are fossil fuels and fissile materials. Examples of environmental effects could include loss of habitat due to dams, loss of habitat due to surface mining, and air pollution from burning of fossil fuels. (NGSS Lead States, 2013a, para 1.)

While a teacher may elect to design instructional activities about wind energy and air pollution, an assessment task his students are administered may address water behind dams and loss of habitat. In making these choices, there are potential consequences for the validity and generalizability of our score interpretations that will be based on whatever choices are selected in the various classrooms.

Flexibility in instructional and assessment contextualization presents an opportunity for the formative use case in which a teacher can select formative assessment tasks that match the characteristics of the instruction—matching context, format, and context-specific content knowledge. When selecting and using a formative assessment task, the teacher can either select a task that uses an identical context to that used in the classroom instruction or, at the very least, consider any differences in the context of the assessment and the instruction when interpreting students' performance. This increases the likelihood that the assessment data will be used to make valid inferences about the skills, ideas, or practices the teacher wanted to assess. In short, this flexibility in the NGSS can strengthen assessment evidence, as well as learning, in instructional, formative, use cases.

The degrees of freedom offered for instruction and assessment that benefit formative use case are more problematic in the summative accountability use case. For large-scale summative accountability assessment, it is impossible for all characteristics of a given assessment task to reflect instruction equally for all students. Further, the inferences we want to make from these tests are about the more broadly defined set of standards, rather than the more focused target of measurement associated with a particular instructional unit. As mentioned earlier, the current strategy for educational assessment is to build tests with multiple discrete items, which tends to wash out any content effects associated with a single task context or content. If, as we argued earlier, MC items are unlikely to capture the hard-to-measure NGSS dimensions of interest, then the MC item solution is not an option. Alternatively, a test with a small number of rich, scenario-based, technology-enhanced tasks, as we have recommended, would be appropriate if it could be established that the three NGSS dimensions are defined in the various performance expectations and are generalizable across any particularities of the specific task. That is to say that the particular context (i.e., format, context-specific content) in which instruction occurs has relatively minor impact with respect to the context in which it is assessed—the skills, practices, and concepts are generalizable to problem solving across tasks. Further, performance in a small number of highly contextualized tasks would be need to be found sufficient to make inferences about the NGSS dimensions that generalize across any context.

Considerable research in cognitive psychology and problem solving suggests that this is not the case (Mayer, 1992; Newell & Simon, 1972; Smith, 1991; Sternberg & Frensch, 1991). Though generalized knowledge that is applicable across domains is useful, context-specific knowledge is often required to

successfully solve a problem. Ruiz-Primo and Shavelson (1996) specifically addressed this issue with respect to performance assessment in science, stating that "whatever performance assessments are measuring about science understanding is sensitive not only to the task and occasion sampled, but also the method used to assess performance" (p. 1051). They concluded that not only would the technical quality of performance assessment require greater scrutiny than brief discrete tasks, but that a large number of performance assessment tasks are needed in order to yield generalizable measures of achievement. Thus, the use of any single contextualized task to assess a given performance expectation could introduce significant bias in measuring individual students. Proficiency might be under or over estimated relative to estimates yielded from a different assessment task. Further, students who have not received instructional tasks with the same specific content knowledge area, context, format, and technology in which it is assessed will be disadvantaged relative to students for whom the instructional and assessment tasks are more closely matched. It would appear, therefore, that the strengths of the highly contextualized tasks which offer promise for capturing the complex NGSS elements could become their biggest liabilities in the summative accountability assessment use case. Moving toward the more complex task types introduces challenges for our repertoire of assessment design practices and tools, a point we will consider next from a more psychometric perspective.

Psychometric Challenges of NGSS Assessment

Psychometric challenges for NGSS assessment are driven by two related factors, the nature of the NGSS themselves and the new task types that they require. The direct impact of the NGSS on psychometric modeling stems from the explicit multidimensionality in the construct as defined by the standards. The indirect impact comes through the characteristics of the tasks required to measure the NGSS. As described previously, it is likely that the tasks needed for formative and accountability NGSS assessment will violate many of the assumptions and capabilities of our traditional psychometric repertoire. Tasks are likely to be scenario based, using a common problem-solving context that creates dependencies between the items. Further, the testing time for any given extended performance task will be lengthy and will limit the total number of tasks a student can be administered. Finally, student responses will generate new data sources, including continuous behavioral data, that do not conform to our traditional models for dichotomous or polytomous item responses. In sum, the psychometric challenges introduced either directly or indirectly by the NGSS include the need for appropriate models to assess dimensionality, to estimate item and person parameters, and then to report out reliable and generalizable scores reflective of status and growth on each of the dimensions targeted.

Dimensionality assessment. A prerequisite to almost all test score scaling and score reporting is the appropriate dimensional specification. That is, in order to properly scale individual students' responses and provide reliable and valid scores to represent status or growth, we must know how many scores should be reported, what they represent, and how they are related to one another. When dimensionality is improperly specified, specifically underspecified, this is both an issue for assessment

development (e.g., when trying to select items to retain or discard) and in scoring. Various approaches to dimensionality have been developed to aid test developers in the design and scoring of test data that maximize reliability and validity.

For the majority of operational educational testing programs, the dimensionality issue would be better characterized as the *unidimensionality* issue, where unidimensionality is the goal and multidimensionality is explicitly avoided. Items that show multidimensionality are removed or revised to yield a test with scores that conform to a unidimensional model. It is unclear whether this has been the tradition because we have typically been interested in measuring unidimensional constructs, or whether we have chosen to focus on unidimensional constructs because we are far more comfortable with unidimensional models and associated tests. Regardless of the motivation, as a result, our psychometric tools to assess dimensionality have been well designed and researched in terms of their ability to identify unidimensional structure or violations thereof (e.g., DETECT; Zhang & Stout, 1999). These tools help us, first, in the development process, to select, remove, or edit items that do not conform to unidimensionality. They can also contribute to our validity argument in providing evidence to justify scaling and reporting on a single unidimensional scale.

Several methods are available to explicitly test multidimensional structure in our high-stakes tests. Among the various methods (e.g., posterior predictive model checking, nonlinear factor analysis, DETECT index), research has shown that under certain conditions, we can detect multidimensionality quite well (Gierl, Leighton, & Tan, 2006; Svetina, 2011). One of the best conditions for detecting multidimensional structure is the case of simple structure, where each item measures one dimension with multidimensionality at the test, rather than the item, level. These procedures are less effective at detecting multidimensional complex structure in which items on a test may measure multiple constructs to varying degrees (Levy & Svetina, 2011; Svetina, 2013). Yet, it is *exactly* the situation that is likely to arise from properly designed measures of the NGSS standards.

The NGSS framework requires that instruction and assessment integrate three types of dimensions—practices, crosscutting concepts, and core ideas. If implemented as prescribed, each standard would integrate one practice, one crosscutting concept, and one core idea into each performance objective, driving both instruction and assessment. NGSS assessment tasks should yield data that reflect all three types of dimensions to greater or lesser extent, each of which is intended to be "reported out on" for either accountability or formative purposes. Consider the example given in the NGSS framework document describing how the three dimensions might be combined based on Organization for Matter and Energy Flow in Organisms (LS1.C), a part of the first core idea in the life sciences. The framework authors suggest that by the end of 5th grade, students should be able to "Explain how animals use food and provide examples and evidence that support each type of use." They detail the criteria for performance in terms of the accuracy and completeness of explanation, with an emphasis on proper argumentation, including use of evidence and diagrams specific to the particular concept in the life sciences. Within this single task, assessment claims include inferences about students'

knowledge of properties of living organisms and their need for energy (disciplinary idea); their ability to develop arguments with claims and evidence (practices); and their understanding of patterns, similarity, and diversity across living things, including matter conservation (crosscutting concepts). The intent is for assessments to use these performance expectations across multiple items to measure the population of grade-relevant standards. The resulting data from such a test comprised of such items are likely to manifest complex structure, with various mixtures of the three dimensions contributing to individual scores from individual test items and tasks. In other words, if properly designed, NGSS tests are likely to generate the conditions that are the most challenging for our repertoire of dimensionality assessment tools. Test development for NGSS tasks will be challenging in so far as assessing the complex dimensional structure is a necessary prerequisite to scaling and reporting.

Scaling and estimation. After an appropriate dimensional solution is reached, there remains the challenge of estimating model parameters for the items and the examinees. Selecting the proper model—compensatory versus noncompensatory, conjunctive versus disjunctive, 1PL versus 2PL versus 3PL—as well as properly estimating the item parameters, can be challenging for most operational testing contexts, not to mention the specific case of NGSS with its complex item formats and all that they entail (i.e., item dependencies, lengthy testing times with few tasks). As compared to their unidimensional counterparts, the multidimensional models can present more complex computational requirements. The increased number of parameters to estimate, including both the individual dimensions and their intercorrelations, place greater requirements on sample size, test length, and distributional characteristics of people and items (Béguin & Glas, 2001; Zhang, 2012). Item parameters, when estimated with insufficient sample sizes or samples with ability distributions poorly matched to the item characteristics, will be difficult to scale. Conversely, items with poorly estimated parameters will lead to inaccurate and unreliable ability estimates for examinees. Recent advances in multidimensional psychometric models, including multidimensional item-response theory (MIRT) models and Bayesian inference networks (BINs), offer promising tools for accurately estimating multiple correlated abilities underlying student performance on complex assessments (Jensen, 2001; Junker & Sijtsma, 2001; Martin & VanLehn, 1995). Further, models designed specifically to account for dependencies arising from common or shared stimuli (e.g., a common reading passage, a shared scenario or context) across items, including testlet models, are available to handle specific sources and structures of multidimensionality on existing assessments (Wainer, Bradlow, & Wang, 2007).

Having the appropriate tools available to model multidimensional data only solves the problem of access to useful models; it does not ensure that they can be leveraged for improved NGSS measurement. A popular approach to psychometric modeling of assessment data is to survey the available models, fit them, compare them, and then select the appropriate model. However, it is our belief that *selecting a model* will not be good enough. What is needed is *building a proper model* from pieces, some of which are familiar structures in familiar models but others which are not, is necessary. Some models (e.g., Bayesian network models, the Generalized Diagnostic Model [GDM], von Davier,

2007) and some statistical software packages (e.g., Mplus, Muthén, & Muthén, 1998; WinBUGS, Lunn, Thomas, Best, & Speigelhalter, 2000) have model-building capabilities. However, building complex tasks and expecting psychometricians to figure out after the fact how to build models is a bad way to proceed. Even differences that seem minor on the surface can have huge modeling implications and evidentiary value implications. The psychometric model-building process must begin much earlier, when we are defining the intended score uses and interpretations in terms of the necessary evidence and the statistical model that will link them. Better to develop a family of inferential structures at the outset around which unique tasks can be developed—so that they (a) can be unique and creative, but (b) for which we know in essence how to score them and how to accumulate evidence up front (Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002). The scoring is designed in, jointly with the substance, and has understood psychometric frames.

Conceding the general challenge of multidimensionality for NGSS assessment, the issue does not necessarily impact all use cases equally. From a statistical perspective, dimensionality is defined in terms of what is needed in a model for particular data to achieve satisfactory approximation to conditional independence. *Dimensionality arises from the interaction of tasks and population, such that the same tasks can be high dimensionality in DIFTS use but low dimensionality when used in instruction.* This point is neither obvious nor unimportant. One hears many very proficient psychometricians and statisticians speak of "the dimensionality of a test," and they do so implicitly assuming a particular population and use case, without qualifying their interpretation. Lord (1976) asserted that dimensionality is a function of instruction and the instructional sensitivity of an item. Thus for two groups of students (i.e., students in different classrooms), one of which received instruction on a set of test items and the others not, test performance on these items that would otherwise have been unidimensional will show additional dimensions. Further, even for a single group of students who have all received the same instruction, the multidimensionality intended in an item may not be evident in the response data. For example, given a set of items designed to measure specific ideas, concepts, and practices, when administered to a population of students who have all mastered the ideas and concepts, the data will conform to a unidimensional model that reflects only variability in practices. This does not make the items any less valid in terms of their reflection of the dimensions listed in the standards—knowledge of the ideas is required to answer the items. It is simply the case that the data can be fit with a unidimensional model due to the nature of the items in relation to the testing population. Thus, if what is intended by the multidimensionality of the NGSS structure is the opportunity to produce scores for each of the three dimensions that could prove diagnostic, it will not be sufficient to consider how the tasks are design, but rather one must understand how the dimensions are reflected in the items relative to the population of test takers and their exposure to instruction.

Given that, in the formative use case, we have considerably more information about what an individual or classroom of students know (or at least what they have been exposed to through instruction), the complex dimensionality that would otherwise emerge in our data may be simplified.

Thus, the dimensionality challenge manifests itself not merely across different populations, but conceivably across different use cases in which population distributions have unique characteristics (e.g., a classroom population for formative assessment vs. a state population for accountability testing). This psychometric inconsistency offers a solution for the dimensionality challenge in the formative case. For accountability assessment, solutions to the psychometric challenges are less clear. While advances have been made in developing appropriate complex psychometric tools through simulation and theoretical research, it has yet to be seen whether such approaches can be supported under the practical constraints of limited testing time, breadth of content coverage, and highly dependent observations like those generated from the recommended NGSS assessment tasks.

While the most recent multidimensional models and estimation algorithms offer the possibility to estimate multiple abilities from a limited set of assessment tasks, there remains the question of whether one *should* attempt to do so. Even when parameter estimates can converge on stable values, the multidimensional inferences about individuals are less efficient than a simpler model. With a limited number of observables (i.e., items or tasks), there is a limited amount of information that can be extracted. It is a consequence of the fact that when multiple unknown factors are driving a relatively small set of observed data, there are multiple competing hypotheses that could explain the observed data. As previously argued, to the extent that we have some local information about individuals' abilities, knowledge, context, and learning experiences, which contribute to their test performance, tasks that are multidimensional *in theory* can be fit with simpler models and provide more useful information on the specific unknown factors. The bottom line is that the multidimensional structure of the construct implied by NGSS can likely be captured via complex psychometric modeling of increasingly complex tasks; however, we should not expect that these models will work like the simple tasks and models used for the more traditional unidimensional standards. What may be more likely is that the complex assessments can provide useful information about trends in learning and thinking for groups of students, rather than for individual students (National Research Council [NRC], 2006). It remains to be seen whether our existing or emerging assessment capabilities can support individual score reports on multidimensional constructs without additional context (i.e., DIFTS assessment).

Reliability and generalizability. Though the issue of score reliability is part and parcel with scaling and estimation, the topic is of sufficient importance that it deserves explicit consideration. We have remarked repeatedly that traditional forced-choice, MC items have dominated high-stakes testing due in part to their conformity to our well-researched psychometric theory. Theories and methods for increasing score reliability and generalizability are certainly among these. In classical test theory, reliability is increased through repeated measurement of the same construct with parallel items for which random error can be averaged out over each individual item response. Simply put, the longer the test, the higher the reliability can be.

Generalizability, a form of reliability, addresses the specific issue of whether, in fact, measurements based on one particular item or test, with its own idiosyncracies and task-specific

characteristics, will remain stable across other items, also with their own specific characteristics. To the extent that any item brings with it particular context, content, or other design features that are not part of the design, increasing the number of items on a test can allow for these factors to have minimal effect on our estimates of the targeted construct(s). Extending the logic of longer tests for increased reliability—longer tests, with more varied contexts and items (all measuring the same construct) will lead to more reliable and generalizable test scores. With constraints on testing time, the strategy of lengthening tests requires that items themselves be relatively short, to allow for administration of large numbers of items.

Modern test theory, specifically item-response theory (IRT), provides psychometric models that can allow for shorter tests to yield comparable reliability to longer ones, under certain conditions. Specifically, if test items are selected to provide maximum information about students by targeting their specific ability level, then fewer items need be administered to achieve targeted levels of precision in our ability estimates. To address one aspect of generalizability, adaptive testing algorithms that leverage IRT models for shorter tests are typically administered with additional specifications regarding the range of item content, format, genres, and so on for any test form. Each student must see a sufficient number of items across the representative content areas and item types on a given test in order for the test scores to be considered valid. Again, allowing for the use of MC or other selected-response item formats allows these requirements to be met without excessive testing times.

Considering these standing practices for increasing score reliability and generalizability, one should become immediately concerned with the compatibility of our recommendations for lengthy, highly contextualized, tasks for NGSS assessment, particularly in the summative/accountability high-stakes use case. For use cases in which priority is placed on generating instructionally relevant information about students' learning and cognition or instructional effectives, as in the interim, formative, and evaluative use cases (Use Cases 3–5), these tasks are necessary for generating appropriate evidence to support the desired use. However, the features of these items that make them so powerful in those use cases are exactly what makes them challenging for reliability and generalizability, which are the priorities in the large-scale use cases (Use Cases 1 and 2). Put simply, the issue is a trade-off between depth and breadth of coverage, both statistical and substantive. With limited testing time, only a small number of the recommended complex assessment tasks would be feasible. Given the dependencies that exist among items associated with a common stimuli, there is concern as to whether sufficient independent pieces of information are provided (*a la* parallel forms) to yield reliability in our score estimates. Even if there were several discrete items within a single complex task, the need to assess multiple dimensions would limit the opportunity to assess any one of them with large numbers of items.

Though not among our two highlighted use cases, complex survey assessments—like NAEP—that target group-level achievement claims may offer the optimal conditions for lengthy assessment tasks that limit the number of items seen by each individual student (see Use Case 2 in Table 1).

Arguably, standards for reliability are relevant only at the level of the target inferences. Thus, for NAEP, the reliability of individual student-level score is of less importance than the reliability and validity of the subgroup means. It is exactly this type of use case–specific variation that should be considered at the outset of assessment design. Use cases in which the cost to reliability resulting from increased complexity in task design provide the greatest opportunity to maximize the benefits of lengthy tasks. The cost to reliability may be too great for the accountability use case in which individual student scores are used for high-stakes decisions. Educators must acknowledge, however, that it is unlikely that we can maximize reliability at the individual score level and maximize the instructional utility and diagnostic value of scores. As the Board on Testing and Assessment (BOTA) report (NRC, 2006, p. 30) points out, three states tried to develop a single assessment to leverage both depth and breadth of coverage. All students were tested on limited subsets of standards, allowing for the use of the complex assessment tasks that would give the depth of information wanted. Every student was tested on some part of the standards, and all standards were measured for at least some students. However, this solution to the complex system demands did not satisfy the customer—public demand and fiscal limitations. The public was not willing to compromise on individual student score reports, and a completely parallel system with both individual and group level targets was not feasible, a point that will be revisited in discussion of logistical and practical challenges to NGSS assessment.

With respect to generalizability, the use of a small set of complex tasks challenges our approach to content and item format sampling to achieve generalizable scores. Additional evidence would likely be required to support the claim that students' behavior in one or two specific contexts is sufficient to make generalizable conclusions about general proficiency on science practices, core ideas, or crosscutting skills. Current research on complex scenario-based assessment suggests that each task[2] could require 30to 45 minutes, which clearly places a limit on the number of independent tasks that can be administered and used to estimate student abilities. Without more information from a broader range of tasks, the possibility of any psychometric model producing generalizable results is low. Thus, while this challenge may have psychometric solutions, it is more likely that what will be needed is creative task design to create generalizable contexts. The challenge will be to identify contexts and tasks that are sufficiently generalizable, but without making them so context-free that they lose their meaning in terms of real-world science problem solving, as mandated by the NGSS framework.

Vertical scaling and construct shift. Our final psychometric challenge relates to our previous discussion of multidimensionality, but with a more targeted focus—multidimensionality when measuring growth due to construct shift. NGSS standards are designed to articulate skill development across grades that should be reflected in the curriculum and associated assessments. This implies a desire to use assessment scores to monitor students' progress as they learn and develop science

[2] *Task* here refers to an intact set of assessment activities that are all associated with a common context, scenario, or other stimulus.

proficiency, as defined by the framework. The NGSS explicitly aim to promote assessments that can be vertically scaled to permit measures of growth over time on the three major dimensions. The appropriateness and feasibility of this goal is simultaneously a psychometric and a substantive one. Vertical scaling is a psychometric approach to statistically linking scores from tests designed to measure the same construct but at different ages or grades. It is the foundation for measuring growth, a key component of most accountability systems for both individual student and teacher performance evaluations. However, the appropriateness of vertical scaling, both as a conceptual process and in practice with complex psychometric models, makes significant assumptions, not the least of which is about the meaning of the underlying construct(s) measured by the vertically equated tests and scores (see Kolen & Brennan, 2004 and Carlson, 2011 for comprehensive discussion of vertical scaling). Specifically, it is assumed that the construct(s) measured by each of the vertically equated tests is properly specified in each of the individual tests, and that the construct(s) is/are *the same* on each measure (Reckase & Martineau, 2004; Wang & Jiao, 2009; Yen, 2007)[3]. The question of construct equivalence is clearly a theoretical one in which we must consider whether, in fact, the nature of a given construct at early grades (i.e., in early development) is psychologically, pedagogically, and practically the same as at older grades and whether the differences constitute, in fact, distinct constructs, as opposed to different locations in terms of sophistication of mental process on the same construct. As a specific case of multidimensionality, construct shift—if ignored—can severely compromise our score interpretations and resulting decisions.

Considerable psychometric research has compared the performance of various linking approaches under a variety of conditions of student ability distributions, test and item characteristics, and dimensional structure (e.g., Kim & Cohen, 2002; Li & Lissitz, 2012; Patz & Yao, 2007; Skaggs & Lissitz, 1988). Most recently, interest has turned to multidimensional and bi-factor IRT models for more accurate linking that leads to more appropriate measures and interpretation of growth than traditional unidimensional constructs. Specific research on multidimensionality of science tests for purposes of vertical scaling has been explored with success in identifying multiple dimensions associated with distinct content under the condition of simple structure (Jiao & Wang, 2008; Reckase & Martineau, 2004; Wang, Jiao, & Severance, 2005). In these applications, the various content areas were modeled as multiple latent dimensions with simple-structure MIRT models, and latent trait correlations were used to determine the strength of the relationships among the dimensions. However, as discussed in the more general issue of multidimensionality, NGSS tasks are likely to give rise to complex, not simple, multidimensional structure. To the extent that we cannot properly model these multiple dimensions, vertically scaled scores may lead to erroneous measures of growth and lead to invalid score interpretations and high-stakes decisions.

[3] Carlson (2001) showed that it is feasible to have a trait that is basically curvilinear but unidimensional within a multidimensional space, and suggested that vertical scaling is one situation where this is likely to occur.

It has been suggested that our best tools for successful vertical scaling may not, in fact, be psychometric in nature. Rather, Briggs (2012) suggests that successful multidimensional vertical scaling hinges on our theoretical definitions of the constructs we want to measure. If we are to properly link scores from tests at different developmental levels, then we should do so with full knowledge of how those developmental levels are related and with items designed to reflect the shifts in purposeful ways. In other words, before trying to statistically link scores, we should have a theoretical justification to do so. Rather than traditional scales and construct definitions comprised of lists of increasingly difficult content, a better conceptual model is offered by more cognitively and developmentally based representations. To this end, learning progressions and other cognitive models of our constructs are invaluable. As compared to traditional construct definitions (i.e., test blueprints and lists of discrete, disconnected performance standards), learning progressions provide cognitively based descriptions of development that directly address our assumptions about construct shift and equivalence (Briggs, 2012).

Consider the conceptual maps of science strands developed by the *Atlas of Science Literacy* Project 2061 (American Association for the Advancement of Science [AAAS], 2001). The maps represent the science domain as a set of interconnected ideas and skills within various content strands and, most importantly, how they build upon one another to lead to science literacy. The content and form of the maps are consistent with the NGSS standards in that they make more conceptual sense based on what we know from the learning sciences. From an assessment perspective, these models provide a strong basis on which to design tasks that can be used to provide better feedback to instruction. However, the structure does not lend itself as well to our traditional scale linking. Further, it is not the case that we expect a student's cognition and performance to be universal across all systems and contexts designed to measure a particular progression. The evidence suggests that people's understanding of systems can vary substantially from one system to another; that individuals' increasing understanding need not follow well-defined levels; and different situations can evoke thinking at different levels described thusly even within the same person (Sikorski & Hammer, 2010). Thus, at best, learning progressions provide us with a theoretical basis upon which we can selectively vertically scale or link scores. We can build tests based on these models and then attempt to link scores and measure growth only across the theoretically and developmentally connected scores. At worst, the result will be a set of tasks yielding scores that may not be vertically linkable, but support more cognitively grounded decisions and interpretations about students' science reasoning and ability.

While it is the case that some learning progressions might be consistent with unidimensional scales, others are built around qualitatively different ways of thinking; that is, construct shift is their essence. Robust learning progressions that articulate distinct ways of thinking will provide a great deal of insight as to how appropriate psychometric models could be built to show change over time. Growth would conceivably be reconceptualized to reflect developments in cognitive theory. Rather than gain scores computed as differences in location on a vertically linked scale from one year to the next, growth would be defined in terms of the number of levels on a learning progression in any particular domain on

which a student has progressed. Models like Wilson's Saltus (1989) that describe development in a stage-like manner might be appropriate for many of these progressions. Still, without a firm grounding on models like learning progressions and items written to measure them, any psychometric procedure is useless. Thus, it is not to say that growth cannot be measured across a complex set of constructs, but it will take time and some trial and error to get it right. Building on past experience with familiar assessments and more recent research gives us a general sense of how to approach the challenge. But how to do so in a particular implementation will take more cycles of piloting and development. If we do not accept this more iterative development path, we risk falling back to what we already know how to do, and limit what can be implemented.

To summarize the psychometric challenges, the basic message is that we know quite well how to model data from our traditional standards-based assessments, because our models were designed for those tasks and data and, conversely, we continue to develop tests based on what we know works for those models. We build tests comprised of similar kinds of tasks and focused mainly on basic concepts and declarative kinds of knowledge. Unidimensional modeling is well understood, and even when assessments are composed of traditional MC items in various multidimensional content domains, they can be handled within the constraints of a unidimensional model through content constraints. However, when the tests are comprised of items that are more complex, such as those that will be produced under NGSS, that simple solution is not available to us. As tasks become more highly dependent on what students study, *and* they press harder on what students are doing with them in terms of inquiry and model-based reasoning, *and* they look for relationships with more abstract overarching ideas, the more the multidimensionality becomes an issue. If you have a test with a variety of tasks assessing wonderfully deep and remarkably broad aspects of science with cutting-edge interactive technology, and drop-it-in on an undifferentiated population, then you get a test for which a unidimensional score has much lower generalizability than old-fashioned tests. For the summative accountability use case in which reliability and generalizability are of most importance, policy makers and educators need to understand what the limits of our current psychometric capabilities are. Simultaneously, psychometricians would be well advised to broaden their perspectives to embrace alternative approaches to statistical modeling and scoring that may not be as familiar or comfortable to us, but are more likely to provide key stakeholders with the type of information that they so desperately need to improve science education.

Logistical and Practical Challenges of NGSS

The majority of the logistical challenges for NGSS, above and beyond those for any large-scale testing program, are consequences of the task design recommendations. As we previously argued, traditional paper-and-pencil multiple-choice tests are the dominant practice because they are feasible and meet the constraints of large-scale assessment contexts. Once we move away from the traditional discrete, paper-and-pencil MC items to technology-based, complex, scenario-based tasks requiring from

30 to 45 minutes each, there are significant implications for access, administration times, security, and cost.

The NGSS framework requires that assessments be administered to all children, in all grades, in all schools. Thus, the cost and logistical requirements of NGSS assessments must be manageable for all districts and schools, regardless of access to technology or budgetary constraints. Clearly, we know that some schools and districts have limited resources and, thus, limited access to technology. With the impending Common Core Consortia assessments set to be released in 2014, all of which require computer-based administration, the question of whether all schools in the United States are equipped with the necessary technology will soon be answered. Regardless of the answer at that particular moment, if recommended NGSS assessments are to be delivered to all students, in all grades, in all schools, there will likely be significant cost associated with bringing the technology to large numbers of schools that do not have the necessary equipment, or whose systems are so out of date that they could not support the innovative technologies we expect to leverage. Memory requirements, processing speeds, graphics cards—all of these must be equally available to every student, classroom, and school, if we are to build appropriate tasks to support NGSS assessment claims. As researchers and policy makers develop idealized plans of what NGSS assessments *could* look like, with the use of games-based and simulation-based tests, we must consider whether it is possible for such a system to be universally inclusive.

Focusing more explicitly on the issue of NGSS for *all children*, issues associated with access for students with disabilities (SWD) may raise challenges. There are two opposing theories about how technology-enhanced complex assessment tasks might impact students with disabilities—it could actually make the assessments more accessible, or it could create new barriers for subgroups of students. The argument for the use of technology-enhanced assessment to increase access for SWD emphasizes the flexibility of the systems and availability of assistive tools (e.g., haptics, refreshable Braille) that are not available with traditional paper-and-pencil tests (Haertel et al., 2012; Laitusis, Buzick, Stone, Hansen, & Hakkinen, 2012; Scalise, 2012). The opposing view posits that some students with SWD will be further challenged by the use of technology, which will only serve to interfere with their ability to demonstrate the targeted KSAs. The likely truth is that the SWD population is so diverse that no one technology will be useful for everyone. The strongest technology is, therefore, the one that can be the most flexible, providing only the accommodations that a given student needs in order to make the task and the response processes more accessible. Ultimately, whether the technology-enhanced tasks in NGSS assessments are a barrier or a source of greater access will depend on the nature of the specific technologies and what affordances they might offer. Test developers and policy makers are advised to undertake the development effort with the explicit intention of testing these assumptions about the technology for SWD and leverage the technological strengths wherever possible.

Beyond the cost of installing appropriate technology in every school for every student, costs associated with complex task development and scoring as recommended here could cause test

development costs to grow significantly. Even with our predominant use of MC items, item development costs are quite high. Large pools of items are needed to ensure item and test security, particularly if adaptive testing systems are used. If similarly large pools of items were required for the types of technology-enhanced tasks we are recommending, the costs could be simply unmanageable. The security issue has been of little concern in the learning context, which is where many of the innovative technologies one might consider using for NGSS assessment were developed (e.g., simulations, games). Item pools for complex performance assessments might arguably need to be even larger than with MC items. Given the rich context and highly engaging activities that would ideally be designed, the tasks are more likely to be memorable to test takers, thus, decreasing the likelihood that any task could be reused or administered at separate testing times within or across years, or even on parallel forms. Finally, even if the funds were available to support the development of large numbers of tasks for multiple forms or adaptive testing, it is not clear that innovative item types will lend themselves well to building large item pools.

Whereas the population of selected response items measuring a single content area may be quite large, the number of complex performance tasks that adequately capture all three of the NGSS targets in realistic contexts may be quite limited. The challenge here will be to see whether we can build efficient item/task generation methods that allow us to use or reuse pieces of the technology across various assessments, forms, and/or tasks. Research on automatic item generation has shown limited success, and even that has focused on multiple-choice items. The use of task models to generate multiple items based on a common set of design features, all of which can be used to support a common set of assessment claims, has been used in some complex assessment contexts (Bejar & Braun, 1999; Frezzo, Behrens, Mislevy, West, & DiCerbo, 2009; Mislevy, Hamel, et al., 2003). These approaches are likely to offer the most promise for scaling in operational use of complex assessment tasks in NGSS.

The final practical challenge, which may be more substantive and psychometric than practical, is the issue of limited testing time. We have just argued that to support the varied use cases for improved NGSS learning, the assessment tasks must be rich, engaging, extended tasks that provide sufficient context for a problem so as to elicit all three of the KSA types intended (i.e., core ideas, practices, and crosscutting skills). Efforts to build rich, scenario-based assessments in other domains have yielded tasks of 15 to 45 minutes, depending on the nature of the targeted standards. As the number of standards to be tested increases, the need for construct and content coverage drives longer and longer testing times. The issue is essentially a depth versus breadth trade-off. If we use extended tasks that give us the rich, cognitively based data we need to support score interpretations and use, we may need to test students for days before we could cover the breadth of standards required for measurement.

This complexity is simplified in certain use cases. In the formative case, teachers are specifically interested in instructionally relevant information about a small number of skills to tailor current instruction. Another use case, the survey assessment use case exemplified by the NAEP program, avoids this challenge through matrix sampling. Even though the NAEP scores must reflect a wide range of

standards, by using complex sampling methods, each student need only provide data on a subset of the total framework. This would allow students to spend the needed time on a complex task that delves deeply into a standard and then aggregate across the students to get adequate breadth on the construct as well. It is unclear how, for the accountability use case in which individual student scores are needed that represent the entire set of standards, complex assessment could be designed with manageable testing time. Again, the solution to one design challenge introduces complexities in another.

Recommended Strategies for NGSS Assessment

Throughout our discussion of the various challenges for NGSS, we have interspersed some recommendations for how to tackle individual challenges. However, the complex assessment design is most daunting when considering the various challenges and potential solutions as a set. As implied throughout our discussion of the individual challenges, solutions to one design priority may come at a cost to others. Further, differences in the specific testing context, purpose, use, and stakeholders will influence the relative importance of each priority, making no single solution appropriate for all assessment designs. In complex educational assessment, there must be design trade-offs with consideration of the specifics of the assessment use case.

Strategy #1: Build a Coherent Assessment System

Of course, the simplest design solution would be to create a common measure to serve all use cases equally well; build a single test that provides summative, formative, and evaluative information. This is appealing both from a cost and time efficiency perspective—reduced cost to the states, reduced testing time, increased instructional time. The intuitive view of assessment design, that one simply designs *a* test, and a test is a test is a test (Braun & Mislevy, 2005) is only tenable under certain conditions. One can get away with such an approach when we have knowledge of the assessment context (e.g., prior learning activities, prior knowledge) and take it into account tacitly in the design, or when the target inference is simple, or the resources are voluminous, or the consequences of errors are negligible. Absent these simplifying assumptions, a more complex solution will require multiple assessments designed to meet specific priorities and goals of a given use case.

Most researchers and assessment developers, including the NGSS framework authors, acknowledge the incompatibilities of test design for each of these uses and have, thus, argued for the need for a complex system of assessment tools, each designed specifically for a particular purpose, but complementary in how they might be used (Bennett, 2010; Bennett & Gitomer, 2009; Herman, 2010; NRC, 2006; Wilson, 2004). The NGSS framework authors call such a view "demonstrably inadequate," stating explicitly that "no single assessment, regardless of how well it might be designed, can possibly meet the range of information needs that operate from the classroom level on up" (NRC, 2012, p. 262). Herman notes that the inherent incompatibility of the varied educational assessment use case goals and requirements severely compromises the extent to which a single test can simultaneously serve each of

the individual purposes. The optimal assessment design is not a particular *universal test*, but rather a coherent assessment system consisting of multiple types of measures, which creates a more comprehensive picture of student learning and serves the needs of the test users across the various use cases (Herman, 2010).

> To achieve successful NGSS assessment system design requires "creative design and comprehensive engineering, moving beyond the current state of the art." Herman (2010) offered a succinct description of the overarching approach we advocate for successful NGSS assessment:
> In a single word but with many steps, I suggest the word "coherence." I believe that by making our assessments more coherent in both design and use, we can create assessment systems which will measure the right stuff in the right ways while better serving intended purposes, particularly the purpose of improving teaching and learning. (p. 1)

Wilson and Berenthal (NRC, 2006) expanded on the principle of assessment system coherence, stating that:

> A successful system of standards-based science assessment is coherent in a variety of ways. It is horizontally coherent: curriculum, instruction, and assessment are all aligned with the standards; target the same goals for learning; and work together to support students' developing science literacy. It is vertically coherent: all levels of the education system—classroom, school, school district, and state—are based on a shared vision of the goals for science education, of the purposes and uses of assessment, and of what constitutes competent performance. The system is also developmentally coherent: it takes into account how students' science understanding develops over time and the scientific content knowledge, abilities, and understanding that are needed for learning to progress at each stage of the process. (p. 24)

We embrace the recommendations of the Board on Science Education (BSE; NRC, 2007) and the BOTA reports (NRC, 2006), which suggest that a coherent assessment system, embedded within the larger educational system, is needed. This would serve to separate the intended assessment uses and desired interpretations, to the extent possible, for which different tests would be designed to support different claims. Separation of assessment tools and uses is needed to distinguish them with respect to uses cases—who needs to know what, for what purpose, with what other information, at what cost and time scale?

Each use case will likely require a different form of assessment, including different sets of tasks and different statistical modeling approaches. The system would include tools to support group-level inferences, others the individual-level claims. Some tools would support feedback loops about instructional decisions, about school performance, maybe even about learning trajectories of individuals. To achieve the goals for one use case, traditional MC tests might continue to serve a critical

purpose. The complex, technology-enhanced tasks may be better suited for others. What is most likely is that some combination of different task types are needed to support the assessment claims of any use case. And by combination of tasks, we do not mean simply task formats, but configurations of tasks with respect to relationships to students' instructional histories, and under different sets of constraints. From a systems perspective, the success of the entire system will result from the coordination of the separate assessments, each with its own distinct purpose, but considered in relation to one another. In designing any one of these tests, compromises will be made that optimize certain priorities at the cost of others. In designing each individual measure, we advocate making design choices purposefully, strategically, and *a priori*.

Strategy #2: Articulate Design Choices

Design trade-offs are nontrivial decisions, where design objectives are conflicting such that no design solution can meet all the objectives completely (Thurston & Nogal, 2001). In these situations, an improvement in one characteristic of the design can only be achieved by limiting another desirable characteristic. The overall goal is design optimization, but how to achieve it is not typically readily apparent. Making a design trade-off is rarely a simple process, and the solution is typically not obvious. As many design problems are complex with no empirically derived algorithm to support the decision-making process, goal prioritization and personal preference are often the basis for design decisions.

Toward this goal, tools that make explicit the design choices and their impact on system outcomes are paramount. In automotive design system optimization, engineers use design structure matrices and axiomatic design to identify the interactions among design features, allowing them to reduce system complicatedness (Ziegler, 2005). These representations serve to highlight and document critical system information that influences design decisions, including system needs, design task sequencing, and design iteration. These tools are intended to reduce development risks, costs, and time by front-loading the design process with careful consideration of the system dependencies. Most importantly, they allow designers to consider consequences of design decisions in early stages of system development when changes can still be made. In the assessment design endeavor, we have a similar tool, evidence-centered design (ECD; Mislevy, 1994; Mislevy, Steinberg, & Almond, 2003). ECD provides a mechanism to separate many of the design decisions one makes in assessment development in order to evaluate their impact on the usefulness and validity of the test scores. The result is increased effort at the outset of the assessment design process to connect more carefully design choices to the intended purpose of the assessment that helps designers identify sources of weakness in the system and suggests ways to improve it to yield scores that are maximally valid for their intended use.

Strategy #3: Borrow Available Information to Simplify NGSS Assessment Complexities

Setting aside the complexities overcome by developing multiple assessments, there are complexities inherent in developing any single NGSS test, most notably the multidimensional, hard-to-measure constructs and all that they imply for psychometrics and task design. These challenges exist for all use cases, but are particularly troublesome for Use Cases 1 and 2, both of which embody DIFTS assessment. To achieve reduced complexity in each of the tests, developers and users should borrow information wherever possible. Borrow information from one test to reduce complexity in another; borrow information of different types of data within any given test; and borrow information from outside of the system, including information about students, teachers, and classrooms. To the extent that the borrowed information can negate alternative hypotheses for observed data that compromise our validity argument, we have simplified the assessment design requirements without losing any of the richness of the assessment score interpretations and uses.

Among the possible use cases, formative assessment (Use Case 4) likely offers the most abundant information that could be considered in conjunction with assessment data. Formative assessment is most useful when it is contextualized with respect to instructional and learning history, as well as indications of where a teacher should take the student next (Heritage, 2008). One of the richest, but often underutilized, sources of data is knowledge from teachers about students' classroom experiences. If we know from information about prior classroom instruction that students have been taught the specific content within the disciplinary core ideas in the same context as a given test item, then it is more likely that a students' ability or inability to answer a question correctly is a function of other dimensions (e.g., practices or crosscutting concepts). Consider the issue of complex dimensionality of NGSS specifically for the formative assessment use case. As suggested in our earlier discussion of dimensionality as a function of person and task interaction, to the extent that we have additional information about the students and their learning and classroom experiences, we can use that information to reduce the number of functional dimensions. When auxiliary information is available about prior learning and instructional activities, as in the formative use case, we can leverage that information to reduce the dimensional space. Therefore, if a teacher wants to assess a student's modeling practices, that teacher should give a task that requires modeling within a core idea that we know the student has mastered, using specific content knowledge areas familiar to the student, and using technology with which the student is experienced.

What about the accountability use case (Use Case 1)? Without this collateral information, only data from the assessment itself can be used to support or refute the targeted score interpretation. In the case when we have little external information about students or the classroom context, we must find strategies to borrow information from within the test itself. One strategy is to lengthen testing time enough that we can gather sufficient data for each test taker on each of the targeted constructs. That is,

borrow more of the same type of information from the same place. Given the complexity of the NGSS standards and the lengthy tasks that are recommended here, this approach would likely place unrealistic strain on the practical design challenges of the system, specifically testing time and testing cost. An alternative is to reduce the number of unknowns in the system by treating some of the parameters as known while estimating others. Bayesian methods use prior estimates for parameters to help with estimation issues. If knowledge about subsets of dimensions or subsets of examinees could be gleaned from other sources—other tests, teacher ratings, knowledge of the current and prior learning—then the dimensionality and number of estimated parameters could be reduced. One aspect of the NGSS framework that may offer some relief in this regard is the requirement that teachers across schools, districts, and states move through the curriculum and the standards at the same pace. Though there would be no way within a DIFTS assessment context to know whether this was the case or not, it might be reasonable to make certain assumptions about the specific learning that has or has not taken place. Further, given the developmental nature of the NGSS, where later, more complex standards are predicated on earlier standards (e.g., prerequisite skills), some assumptions could be made about student processing, thus, allowing for certain parameters to be fixed and treated as known. The result would be assessments that reflect the desired multidimensionality of the NGSS in their design, but pose a more unidimensional (or at least lower dimensional) problem for statistical estimation.

An alternative strategy for borrowing information for the accountability use case is to borrow information from different data sources within the same test. We have stated that our traditional MC item types have desirable psychometric properties in that they work well with our traditional models, but are limited in their instructional and diagnostic utility; conversely, complex assessment tasks offer a great deal of rich cognitive information, but may not be so easily modeled. The strategy is to borrow strength from each—the psychometric stability of our traditional data and the cognitive richness of the new data sources. Scalise (in press, 2013) developed a hybrid MIRT-Bayes model that accomplishes exactly this goal. In modeling assessment data from a complex science learning environment, she collected both traditional assessment response data and log-file data that captured students' interactions with technology-delivered tasks. Using the hybrid model, the measurement precision and construct relevance of the scores were both enhanced by simultaneously modeling how students reasoned about the science content, approached the problem-solving process, and generated solutions to the task. Innovative psychometric models, like the hybrid MIRT-Bayes model, when applied to appropriately designed tasks that elicit construct-relevant evidence of the NGSS, are likely to offer the most promising solutions for our large-scale accountability assessment challenges.

Conclusion

With the goal of improving the quality of science education, the NGSS release has fundamentally changed the nature of our science assessment goals. All of the challenges that we have discussed in this paper are consequences of that initial increased complexity about what we want our NGSS assessments

to be able to do. As compared to our traditional sciences assessments, we want our NGSS assessment scores to support interpretations about multiple hard-to-measure constructs and do so in a way that can both support high-stakes decisions and be diagnostic and instructionally relevant. Given that much of our discussion has focused on the challenges of the NGSS for assessment, the reader may interpret our view of the added complexity as negative. To be clear, we believe that the effort required to overcome the challenges is well worth the cost. Science instruction and assessment must be reflect the complexities in modern views of science, specifically the need to focus simultaneously on the blending of practices, ideas, and concepts, or we will fail our students in preparing them to be effective scientific contributors to the global workforce. It will take some work to lay out desired use cases and to derive component assessments with task designs, scoring methods, and psychometric models that satisfy the new requirements appropriately for each assessment purpose. Trying to build a system without doing this work makes failure more likely and wasted resources certain.

What we hope to have shown through our discussion is that the NGSS assessment system, as any complex design system, will require trade-offs. We should simplify the complexities to the extent that we can, either by borrowing information from other sources to reduce system demands, by explicitly making compromises in our demands on the system, or by researching alternative methods that achieve the results of the complex techniques but through simpler means. These are all strategies that have served other complex system design endeavors well, but the relative youth of our *science* as compared to other disciplines limits what can be done. As policy makers and educators push for complex standards, like the NGSS, they must be aware of the current state of the art in assessment, which may not be able to accomplish all that they would like to believe. Continued research on complex assessment design, including research on learning progressions and other cognitive models, innovations in technology-enhanced complex task design, and multidimensional psychometric modeling offer the most promising directions to support optimal NGSS assessment system. As advances in these areas are made, however, it would behoove all stakeholders in the education context—teachers, students, parents, policy makers, assessment developers, administrators—to communicate with one another about each of our needs, the desired goals of the system, and the reality of our current capabilities. For without communication, there can be no coherence in our assessment systems as they serve to improve educational outcomes and attainment for all students.

Author Note

Any opinions expressed in this paper are those of the author(s) and not necessarily of Educational Testing Service.

References

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. Washington, DC: Author.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics, 20,* 95–109.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 471–488.

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). An evidence centered design for learning and assessment in the digital world. In M. C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–54). Charlotte, NC: Information Age.

Bejar, I. I., & Braun, H. (1999). *Architectural simulations: From research to implementation: Final report to the National Council of Architectural Registration Boards* (Research Memorandum No. RM-99-02). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8,* 70–91.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York, NY: Springer.

Braun, H. I., & Mislevy, R. J. (2005). Intuitive test theory. *Phi Delta Kappan, 86,* 488–497.

Briggs, D. C. (2012, April). *Making inferences about growth and value-added: Design issues for the PARCC consortium.* Paper presented at the meeting of the National Council on Educational Measurement, Vancouver, Canada.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006, April). *Developmental assessment with ordered multiple-choice items*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Carlson, J. E. (2001, April). *Curvilinear dimensions of tests and items*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Carlson, J. E. (2011). Statistical models for vertical linking. In A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 59–70). New York, NY: Springer.

Carnegie Corporation of New York & Institute for Advanced Study. (2007). *The opportunity equation: Transforming mathematics and science education for citizenship and the global economy.* Retrieved from http://opportunityequation.org/uploads/files/oe_report.pdf

Champagne, A. B., Kouba, V. L., & Hurley, M. (2000). Assessing inquiry. In J. Minstrell & E. H. Van Zee (Eds.), *Inquiring into inquiry learning and teaching in science* (pp. 447–470). Washington, DC: American Association for the Advancement of Science.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (Research Report No. RR-63). New York, NY: Consortium for Policy Research in Education.

diCerbo, K. E., & Behrens, J. T. (2011). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 13–55). Charlotte, NC: Information Age Press.

Frezzo, D. C., Behrens, J. T, Mislevy, R. J., West, P., & DiCerbo, K. E. (2009). Psychometric and evidentiary approaches to simulation assessment in Packet Tracer software. In *ICNS '09: Proceedings of the fifth international conference on networking and services* (pp. 555–560). Washington, DC: IEEE Computer Society.

Gierl, M. J., Leighton, J. P., & Tan, X. (2006). Evaluating DETECT classification accuracy and consistency when data display complex structure. *Journal of Educational Measurement*, *43,* 265–289.

Gorin, J. S. (2006). Item design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4)*,* 21–35.

Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation theory and practice* (pp. 136–156). New York, NY: Taylor & Francis.

Gorin, J. S., & Svetina, D. (2011). Test design with higher order cognition in mind. In G. Schraw & D. H. Robinson (Eds.), *Assessment of higher order thinking skills*. Charlotte, NC: Information Age Publishing.

Green, B. (1978). In defense of measurement. *American Psychologist, 33,* 664–670.

Haertel, G. D., Cheng, B. H., Cameto, R., Fujii, R., Sanford, C., Rutstein, D., & Morrison, K. (2012, May). *Design and development of technology enhanced assessment tasks: Integrating evidence-centered design and universal design for learning frameworks to assess hard-to-measure science constructs and increase student accessibility.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Educational Testing Service, Washington, DC. Retrieved from http://www.k12center.org/events/research_meetings/tea.html

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). New York, NY: Routledge Taylor and Francis Group.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do*? Phi Delta Kappan, 89,* 140–145.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment.* Washington, DC: Council of Chief State School Officers.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers.

Herman, J. L. (2010). *Coherence: Key to next generation assessment success* (AACC Report). Los Angeles: University of California.

Jensen, F. V. (2001). *Bayesian networks and decision graphs.* New York, NY: IEEE Computer Society.

Jiao, H., & Wang, S. (2008, April). *Construct equivalence for vertically scaled science assessment*. Paper presented at the meeting of the American Educational Research Association, New York, NY.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational and Behavioral Statistics*, *20,* 175–190.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26,* 25–41.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer Science + Business Media.

Laitusis, C., Buzick, H., Stone, E., Hansen, E., & Hakkinen, M. (2012, June). *Literature review of testing accommodations and accessibility tools for students with disabilities*. Paper presented at the Smarter Balanced Assessment Consortium, Princeton, NJ. Retrieved from http://www.ous.edu/sites/default/files/dept/k16align/SBStudentsWithDisabilitiesLitReview.pdf

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models.* Cambridge, UK: Cambridge University Press.

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of mathematical and Statistical Psychology, 64,* 208–232.

Li, Y., & Lissitz, R. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36,* 3–20.

Lord, F. M. (1976). *A study of item bias using characteristic curve theory*. Retrieved from ERIC database. (ED137486)

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337.

Martin, J., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. L. Brennan (Eds.). *Cognitively diagnostic assessment* (pp. 141–165). Hillsdale, NJ: Erlbaum.

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York, NY: Freeman.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J., Hamel, L., Fried, R. G., Gaffney, T., Haertel, G., Hafter, A., … Wenk, A. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report No. 1). Menlo Park, CA: SRI International. Retrieved from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf

Mislevy, R. J., Johnson, E. J., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 20,* 131–154.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15,* 363–389.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

National Research Council. (2006). *Systems for state science assessment.* Washington, DC: National Academies Press.

National Research Council. (2007). *Taking science to school*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: National Academies Press.

NGSS Lead States. (2013a). *4-ESS3-1 Earth and human activity.* Retrieved from http://www.nextgenscience.org/4-ess3-1-earth-and-human-activity

NGSS Lead States. (2013b). *How to read the Next Generation Science Standards (NGSS).* Retrieved from http://www.nextgenscience.org/sites/ngss/files/How%20to%20Read%20NGSS%20-%20Final%204-19-13.pdf

Newell, A., & Simon, H. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253–272). New York, NY: Springer Science + Business Media.

Quellmalz, E. S., & Haertel, G. D. (2004). *Use of technology-supported tools for large-scale science assessment: Implications for assessment practice and policy at the state level.* Commissioned paper prepared for the National Research Council's Committee on Test Design for K–12 Science Achievement, Washington, DC.

Reckase, M., & Martineau, J. (2004). *The vertical scaling of science achievement tests.* Commissioned paper prepared for the National Research Council's Committee on Test Design for K–12 Science Achievement, Washington, DC.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*, 1045-1063.

Rupp, A. A., diCerbo, K. E., Levy, R., Benson, M., Sweet, S., Crawford, A., … Behrens, J. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4,* 49–110.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of episiemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8,* 1–47.

Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4,* 1–10.

Rust, K. F., & Johnson, E. G. (1992). Sampling and weighting in the National Assessment. *Journal of Educational and Behavioral Statistics, 20,* 111–129.

Scalise, K. (in press). Hybrid measurement models for technology-enhanced assessments through mIRTBayes. *Psychometrika*.

Scalise, K. (2012, May). *Using technology to assess hard-to-measure constructs in the CCSS and to expand accessibility.* Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, Educational Testing Service, Washington, DC. Retrieved from http://www.k12center.org/events/research_meetings/tea.html

Scalise, K. (2013). *Multiple grain sizes of inference in innovative assessments: mIRT-Bayes as a hybrid measurement model.* Paper presented at the Center for Educational Assessment, University of Massachusetts, Amherst.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment, 4*. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653/1495

Schmidt, W., Houang, R., & Cogan, L. (2002). A coherent curriculum: A case of mathematics. *American Educator, 26,* 10–26.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.). *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age Publishers.

Sikorski, T., & Hammer, D. (2010). A critique of how learning progressions research conceptualizes sophistication and progress. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 2010 International Conference of the Learning Sciences* (pp. 277–284). Chicago, IL: ISLS.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 14,* 23–32.

Smith, M. U. (1991). A view from biology. In M.U. Smith (Ed.). *Toward a unified theory of problem solving* (pp. 21–34). Hillsdale, NJ: Erlbaum.

Sternberg, R. J., & Frensch, P. A. (1991). *Complex problem solving: Principles and mechanisms.* Hillsdale, NJ: Erlbaum.

Svetina, D. (2011). *Assessing dimensionality in complex data structures: A performance comparison of DETECT and NOHARM procedures* (Unpublished doctoral dissertation). Arizona State University, Phoenix, AZ.

Svetina, D. (2013). Assessing dimensionality in noncompensatory MIRT with complex structure. *Educational and Psychological Measurement, 73,* 312–338.

Thurston, D. L., & Nogal, A. (2001). Meta-level strategies for reformulation of evaluation function during iterative design. *Journal of Engineering Design*, *12,* 93–115.

von Davier, M. (2007). *Mixture general diagnostic models* (Research Report No. RR-07-32). Princeton, NJ: Educational Testing Service. Retrieved from http://www.em2007.mpg.de/files/RR-07-32.pdf

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York, NY: Cambridge University Press.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69,* 760–777.

Wang, S., Jiao, H., & Severance, N. (2005, April). *An investigation of growth patterns of student achievement using unidimensional and multidimensional vertical scale methods.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing.* Mahwah, NJ: Erlbaum.

Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.

Wilson, M. R. (2004). *Towards coherence between classroom assessment and accountability*. Chicago, IL: University of Chicago Press/The National Society for the Study of Education.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational and Behavioral Statistics, 20,* 155–173.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.). *Linking and aligning scores and scales* (pp. 273–283). New York, NY: Springer.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36,* 375–398.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 321–249.

Ziegler, R. J. (2005). *Complexity reduction in automotive design and development* (Unpublished master's thesis). Massachusetts Institute of Technology, Boston, MA.

The Center for K–12 Assessment & Performance Management at ETS creates timely events where conversations regarding new assessment challenges can take place and publishes and disseminates the best thinking and research on the range of measurement issues facing national, state, and local decision makers.

K-12 Center at ETS

Invitational Research Symposium on
Science Assessment