Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores (SCILLSS)

The Role of Performance Level Descriptors for Establishing Meaningful

and Useful Reporting Scales in a Principled Design Approach

Richard M. Luecht

January 2019

The Role of Performance Level Descriptors for Establishing Meaningful and Useful Reporting Scales in a Principled Design Approach was developed with funding from the U.S. Department of Education under Enhanced Assessment Grants Program CFDA 84.368A. The contents do not necessarily represent the policy of the U.S. Department of Education, and no assumption of endorsement by the Federal government should be made.

Table of Contents

List of Exhibits

Section 1.0: Introduction

Performance level descriptors[1] (PLDs) are used for many educational assessments as a means of describing expectations about students' levels of knowledge and skills. Historically, PLDs have been applied as part of a standard setting process to formally articulate the threshold performance expectations associated with a test's cut scores for classifying students into categories such as *basic*, *proficient,* and *advanced*. Unfortunately, there is often a serious disconnect between the broad language of PLDs and the defensible, evidence-based meaning of "successful" performance along a score scale[2].

This white paper provides an overview of PLDs and discusses some of the issues and challenges encountered when developing and implementing useful PLDs in practice. The paper goes on to argue that the *Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores* (SCILLSS) project presents a unique opportunity for its partner states to take advantage of modern evidence-based, principled assessment design strategies and procedures to develop PLDs that are useful and demonstrate strong alignment between task-focused performance expectations and student test scores.

This proposed evidence-based, design-oriented approach to PLD development is not new, but it does deviate substantially from traditional approaches to test development, psychometric analysis, and standard setting procedures. Simply put, the proposed approach *builds validity-focused performance expectations into the upfront design of every test item and carries those expectations forward as concrete interpretations of student scores*. Making this happen in practice is not simple—which is why some testing organizations may opt for traditional approaches. However, it is the right thing to do from the perspective of creating fair, useful, and truly valid test-score interpretations.

Section 1.1: Some Upfront Recommendations

Assessment design and implementation is a complex process. PLD development is an important part of that process. To ensure an effective, integrated PLD design strategy, the following four recommendations, as described in reference to three-dimensional science standards (e.g., the Next Generation Science Standards (NGSS)), should be used: (1) envisioning of the score scale(s) from an evidence-based design perspective; (2) articulating the PLDs relative to each reporting scale; (3) mapping the test content (e.g., NGSS content knowledge and skills specifications) to the reporting scale(s) and PLDs using item and task designs that meet the evidence requirements along the reporting scale; and (4) committing to prospective, iterative design principles. These recommendations are briefly discussed below and addressed in greater depth throughout this white paper.

Envisioning the Score Scales

SCILLSS partner states, working with their vendors, need to create a concrete, policy-directed vision of the intended score scale(s). This vision should specify the number of scales and include a well-articulated progression of performance claims and evidence requirements across each reporting scale. This critical first step is not trivial; to be useful, the ensuing PLDs need to be aligned to the proficiency claims and evidence requirements for each score scale—they cannot be generically stated in broad language that crosses grades and different score scales. In the context of the NGSS, SCILLSS partner states could elect to have one overall reporting scale (per grade) that integrates the three NGSS

[1] PLDs are synonymously called *achievement level descriptors* (ALDs) in quoted sources in this paper.
[2] The terms score scale and reporting scale will be used interchangeably throughout the paper.

dimensions: disciplinary core ideas (DCIs), cross-cutting concepts (CCCs), and science and engineering practices (SEPs). States could also elect to have multiple reporting scales. For example, states could envision three reporting scales, each corresponding to one of the NGSS dimensions (DCIs, CCCs, or SEPs). Or, states could decide to go with four disciplinary score scales (physical sciences, life sciences, Earth and space sciences, and engineering and technology).

There are rather obvious cost and resource implications associated with developing more than one reporting scale. However, those economic implications should be seriously considered relative to a trade-off with the specificity of the content knowledge and proficiency interpretations we can make from the scores. More score scales provide more interpretive specificity about what students know and can do (e.g., to potentially guide instruction and learning). However, building and maintaining multiple score scales over time can significantly increase costs[3]. Germane to the present discussion, simply ignoring the number and nature of the score scales envisioned for the SCILLSS project until some later date will severely curtail the utility of the scores and PLDs and, even more seriously, could lead to a somewhat incoherent design that is little different than traditional assessment design.

Articulating the PLDs for Each Score Scale

To be useful, the PLDs should define the concrete performance expectations relative to a score scale. The NGSS grade-level performance expectations provide guidance but are not detailed enough nor articulated in language that is amenable to what can typically be measured on a test, and by extension, the inferences that can be drawn from the test scores. The PLDs can provide the "connective tissue" that links grade-specific, measurable performance expectations to the proficiency claims and evidence requirements envisioned for each score scale. Arguably, the utility of generic PLDs is extremely limited (speaking here of PLDs that use broad language and performance/achievement descriptors that reflect content knowledge or skills not necessarily measured by a specific test). Furthermore, rather than waiting until the score scale is developed and then determining the PLDs—typically through an appropriate standard setting process—this paper recommends: (1) determining the number of PLDs for each reporting scale, and (2) developing and integrating the PLDs relative to the progression of proficiency claims and evidence requirements for each reporting scale. These initial PLDs can certainly be clarified (i.e., slightly modified) as future versions once the assessments become operational. However, it is essential to have those performance expectations reflected in the assessment design (i.e., mirroring the proficiency claims and evidence requirements along a score scale).

Mapping the Test Content Specifications to the Score Scale and PLDs

Once the score scale(s) requirements are specified in terms of evidence-based proficiency claims and the PLDs are initially specified for each score scale, we can begin creating item and task design specifications that reflect the evidence-based task requirements along the score scale—possibly prioritizing the test content to optimize the precision of PLD-based student proficiency classifications. That is, we want the test content specifications to include concrete design models that directly relate the test item and task content-based knowledge and skills to the evidence requirements along the score scale. The alignment between the envisioned score scale(s), the PLDs, test form content, and the actual test scores therefore becomes concrete *by design*. Harris, Krajcik, Pellegrino, & McElhaney (2016) provide one example for generating this type of alignment in the context of NGSS-based classroom assessments.

[3] There are ways to engineer the test and item production systems to provide scalable solutions that reduce costs and resource requirements over time. However, those procedures and systems must likewise be integrated in the over design plan.

Committing to Iterative Design

One of the sometimes painful lessons in the software design and industrial engineering design fields has been that few initial designs are perfect. For example, "version 1.0" of most software programs will have unanticipated bugs that are fixed in subsequent versions. The educational testing industry is no different. Iterative design requires a commitment to refinement and improvement over time as well as a realistic design strategy that plans for change in terms of short- and longer-term goals. From a policy perspective, this commitment can somewhat mitigate unwarranted and unrealistic expectations for perfection from the onset. But it also ideally allows for appropriate funding and resource allocations to ensure that future versions can actually materialize to resolve or improve the efficiency of initial design flaws.

Section 1.2: Some Background and Perspectives on PLDs

PLDs have an extended history of use in education. For example, the National Assessment of Educational Progress (NAEP) has three PLDs labeled *Basic*, *Proficient,* and *Advanced*. Educational policy makers usually determine the number of PLDs and may further adopt general policy definitions about the level of performance or rigor implied by each PLD, regardless of grade or content domain. Continuing the NAEP example, "Students performing at or above the Proficient level on NAEP assessments demonstrate solid academic performance and competency over challenging subject matter…" (NCES, 2012). These types of policy definitions help ensure a consistent interpretation of the rigor and performance expectations within state testing programs or other educational assessment settings like the NAEP.

A systematic development process is undertaken to refine the PLD definitions for specific grades and for the content. For example, the three PLDs for the NAEP Grade 8 Science assessment are as follows.

> **Basic.** Students performing at the Basic level should be able to state or recognize correct science principles. They should be able to explain and predict observations of natural phenomena at multiple scales, from microscopic to global. They should be able to describe properties and common physical and chemical changes in materials; describe changes in potential and kinetic energy of moving objects; describe levels of organization of living systems—cells, multicellular organisms, and ecosystems; identify related organisms based on hereditary traits; describe a model of the solar system; and describe the processes of the water cycle. They should be able to design observational and experimental investigations employing appropriate tools for measuring variables. They should be able to propose and critique the scientific validity of alternative individual and local community responses to design problems.
>
> **Proficient.** Students performing at the Proficient level should be able to demonstrate relationships among closely related science principles. They should be able to identify evidence of chemical changes; explain and predict motions of objects using position time graphs; explain metabolism, growth, and reproduction in cells, organisms, and ecosystems; use observations of the Sun, Earth, and Moon to explain visible motions in the sky; and predict surface and ground water movements in different regions of the world. They should be able to explain and predict observations of phenomena at multiple scales, from microscopic to macroscopic and local to global, and to suggest examples of observations that illustrate a science principle. They should be able to use evidence from investigations in arguments that accept, revise, or reject scientific models. They should be able to use scientific criteria to propose and critique alternative individual and local community responses to design problems.

> **Advanced.** Students performing at the Advanced level should be able to develop alternative representations of science principles and explanations of observations. They should be able to use information from the periodic table to compare families of elements; explain changes of state in terms of energy flow; trace matter and energy through living systems at multiple scales; predict changes in populations through natural selection and reproduction; use lithospheric plate movement to explain geological phenomena; and identify relationships among regional weather and atmospheric and ocean circulation patterns. They should be able to design and critique investigations involving sampling processes, data quality review processes, and control of variables. They should be able to propose and critique alternative solutions that reflect science-based trade-offs for addressing local and regional problems. (NCES, 2012)

Students are classified into these performance level categories by comparing their test scores to cut scores that have been determined through a formal standard setting process (Loomis & Bourque, 2001; Hambleton, 2001). In this sense, the PLDs often define the *threshold* or *borderline* performance characteristics that distinguish students at adjacent levels (Perie, 2008). This subtle distinction about threshold performance at the cut score implies that PLDs do not necessarily describe what students should know or be able to do across the entire performance level category. The PLDs may also exclude the lowest level of performance since students not falling into one of the other categories are automatically classified as being "below" the lowest category. For example, NAEP does have an implied "Below Basic" category.

In this brief introduction, it is worth highlighting three important aspects about traditional PLD development. First, there is an implied top-down development strategy that starts with fairly broad policy statements about the rigor and expected levels of performance across the assessment enterprise (e.g., across all state end-of-course and end-of-course examinations). A refinement process is then applied to develop more detailed PLDs within grade-specific content domains—refinement that may extend to specific grade-level standards within each domain and grade. Finally, borderline performance expectations are integrated in an approved standard setting process that determines the final set of cut scores—the performance standards. Second, most performance expectations ignore or at least downplay the notion of *task complexity*—that is, the level of and type(s) of challenge(s) presented to the student as (s)he progresses from one proficiency level to the next. Third, there is an implication that PLDs, being integral to standard setting, should primarily describe borderline performance, rather than defining performance expectations across regions of the score scale.

These aspects of PLD development are important to address in this white paper to support the argument for an assessment design process different from a traditional approach. While traditional PLD development can provide some guidance on test development, there is a direct utility in developing PLDs from a "bottom up" perspective that more directly aligns the specific skills, content, and cognitive complexity of tasks used to construct tests and inform interpretations of the score scale, and then, if needed, generalize to some higher level of aggregated, policy-based statements about the rigor of performance expectations in a particular domain. It is further reasoned that there is enormous merit in considering differential task and content *complexity* as an integral part of the PLD definitions. For example, how might scaffolding such as leading questions or guidance tools be more directly integrated into the PLDs at the lower proficiency levels? Finally, an over-arching assumption is that the PLDs need to reflect performance expectations across the entire score scale, not just at implied borderlines between proficiency-level categories.

Section 1.3: Organization of this White Paper

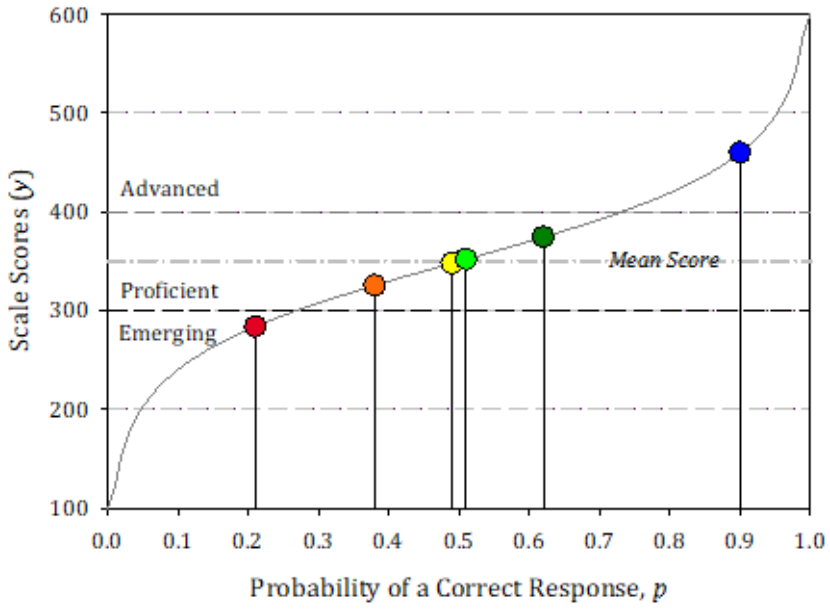
The recommendations offered in Section 1.1 permeate all the sections of this paper and are further elaborated specifically for the context of designing NGSS-based measurement scales, PLDs, and assessment item and test specifications for the SCILLSS partners. Section 2.0 discusses the practical challenges of score interpretation and the implications of those challenges for developing PLDs. Section 3.0 discusses various perspectives on and strategies for developing PLDs, including a multilevel framework proposed by Egan, Schneider, and Ferrara (2012) and implemented by the Smarter Balanced Assessment Consortium (SBAC). Section 4.0 presents a principled assessment design perspective and strategy that can overcome some of the short-comings of more the traditional PLD-development strategies outlined in Section 3.0. NGSS-relevant examples are included. Finally, Section 5.0 offers some recommendations moving forward.

Section 2.0: On the Meaningfulness of Scores

The first recommendation in Section 1.1 suggested that the SCILLSS partners first focus on envisioning one or more score scales. Why is that important when discussing PLD development? An appropriate answer is that, since PLDs describe performance expectations that are operationalized as cut scores on a score scale, it would seem logical that they should be useful for interpreting individual student scores. However, even with well-developed PLDs, there remains a serious challenge related to meaning and use of the scores (Messick, 1989, 1994; Kane, 2006; Zenisky, Hambleton, & Sireci, 2009). A relatively simple example illustrates this score scale interpretation problem.

Exhibit 1 shows a score scale along the vertical (*y*) axis that ranges (bottom to top) from 100 to 600 points. That range of points is fairly typical for state end-of-grade and end-of-course tests. The horizontal axis (*p*) denotes the probability of an examinee correctly responding to a question of average difficulty on the test. The curve in Exhibit 1 implies that the probability of a correct response increases as a function of higher scores. For example, at a score of y=400, the probability of a correct response to the average question on this test is *p*=0.73.

Exhibit 1. Scale Scores and the Probability Correct to an Average Test Question



Now, let's provide a bit more meaning for this score scale. First, the center of the scale score (y=350) corresponds to *the average scale score of all students in an unnamed state who were administered a mathematics test for the first time, last year*. It also corresponds to a 50:50 chance of correctly responding to a typical test item on the scale (i.e., p=0.5 when y=350). Second, scores above 400 are classified as "advanced" performance, scores between 300 and 400 are labeled as "proficient," and scores below 300 are categorized as "emerging." As discussed further on, these classifications could be considered PLDs. If we look closely, we can further surmise that students scoring at the "advanced" level will tend to get approximately 75 percent of the items correct on this test. Conversely, students scoring in the "emerging" category are likely to get, at most, about 25 percent of the items correct. Is that information meaningful? How is it meaningful? To whom is it meaningful?

We might next contemplate the six scores denoted by the colored dots in Exhibit 1. The lowest score (red dot) corresponds a scale score of y_1=284. We can surmise the following about this score. It is below the mean and falls within the "emerging" proficiency category. Students with this score would only have about a 20 percent chance of correctly answering a typical question on this test (or, more exactly, p_1=0.21). What about the four scores within the "proficient" category (orange through green dots)? Two of the scores are below the mean (y_2=326, p_2=0.38, y_3=348, and p_3=0.49), and two scores are above the mean (y_4=352, p_4=0.51, y_5=374, and p_5=0.62). How different are those scores? The scores of y_3=348 and y_4=352 are only four points apart and the associated probability values (p_3=0.49 and p_4=0.51) a nearly identical chance of correctly responding to the items. However, when we reconsider that one (yellow dot) is "below the mean" and the other (green dot) is "above the mean," that qualitative difference may seem important from an interpretative perspective. Finally, consider the highest score of y_6=460 (blue dot) with a probability of p_6=0.9. This "advanced" student is clearly performing above the mean at a higher level than the students with lower scores and seems likely to correctly answer almost all of the questions on the test.

How much more content does the highest-scoring student know relative to the next highest scoring student? If we say, "Well, (s)he knows 90% of the content on the test," is that meaningful? What is the content? What are the test questions? How are they scored? In each of these cases, whether we use the scale scores, the probabilities, or the proficiency categories, there is deficit of meaning as to the nature and extent of the differences relative to the knowledge and skills that are supposed to be measured by the test. We might show the test-content blueprint and even map some prototypical items to the score scale based on their statistical difficulty. But, that would not add much meaning since most content blueprints are undifferentiated with respect to the difficulty of the test questions and prototypical items are limited in their generalizability. *This becomes the principal challenge addressed in this paper—to add meaningful interpretation to the score scales!*

As broad policy definitions, PLDs may be useful to some as a means of articulating the expected student performance and rigor for a testing program. However, most PLDs could be stated without any reference to a specific test or set of scores. A simple exercise may help make that point. Suppose that, after reading the NAEP Grade 8 Science PLDs described earlier, we are told that the corresponding cut scores are 141 (Basic), 170 (Proficient), and 215 (Advanced). Do the NAEP PLDs make those numbers meaningful and, if so, in what ways would that prove useful to teachers, students, and parents?

It seems obvious that PLDs are supposed to be an integral part of the interpretive argument about the meaning and use of scores (Cizek & Bunch, 2007; Perie, 2008; Zenisky, et al, 2009). But, how useful are PLDs for actually interpreting scores? What evidence do we have that a student with a given score can

consistently perform as implied by the corresponding PLD? Huff and Plake (2010) offered a succinct answer to that question:

> Regardless of which approach one takes to constructing a validity argument, it will essentially require that intended interpretations about student performance are supported by evidence; however, despite the accumulation of this validity evidence, it often remains unclear exactly what a student who is classified as proficient is expected to know and be able to do (Linn, 2005; Perie, 2008). When our descriptions of what we expect students to know and be able to do (i.e., ALDs) are created in the absence of sound methodology, are ambiguous, or are not explicit with respect to their influence on the tests that are used to assess students, then a significant component is missing from our validation argument. (p. 131)

The consensus opinion appears to be that PLDs should be developed as early as possible in the test design and development process (AERA/APA/NCME, 2014; Bejar, Braun & Tannenbaum, 2007; Hambleton & Pitoniak, 2006; Perie, 2008; Zieky, Livingston, & Perie, 2008). However, this is often easier said than done and there is very little research as to effective and best practices for using PLDs to drive item and test design and development. In many cases, the test specifications (i.e., the content *blueprints*) are developed without considering the psychometric properties and interpretations of the score scale. Furthermore, given the vague details and guidance provided by most content blueprints, item writers may redefine the nature of the construct being measured when they create new test items. In turn, the test forms are developed by choosing items that happen to correlate well with each other, while possibly meeting item difficulty targets and content constraints. Ultimately, the measurement properties of the score scale rely on the quality of the items on each test form. Sophisticated psychometric analyses may then be used to derive the actual scores[4]. Finally, PLD and standard setting practices are layered onto the score scale to ideally provide meaning for the numbers.

Section 3.0: Traditional Perspectives and Strategies for Developing PLDs

Although some groups wait until cut scores are set to write the PLDs, most guidelines recommend generating PLDs as early as possible in the standard setting process (Loomis & Bourque, 2001; Bejar et al, 2007; Perie, 2008; Huff & Plake, 2010; Plake, Huff, & Reshetar, 2010). The typical development steps are to: (i) specify the number of performance levels as a matter of policy[5]; (ii) draft the initial policy definitions with input from appropriate constituencies; (iii) elaborate/detail the policy definitions for each grade and subject; and (iv) refine the grade- and subject-specific descriptions based on standard setting—including possible "vertical articulation" across grades and subjects.

Unfortunately, the rather exclusive focus on policy-based PLDs can create an enormous interpretation gap. That is, broad policy-based PLDs, even when crafted within the language of specific grades and subjects may seem to be related to a score scale via their association with the cut scores established by the standard setting process. However, the usefulness of those interpretations in terms of what the score scale means is still nebulous. This interpretive gap suggests that perhaps there is a need for different types of PLDs.

In an attempt to better clarify the differential roles of PLDs, Egan, Schneider, and Ferrara (2012) suggested that there may be four types of PLDs: (a) Policy PLDs as general descriptors that articulate the

[4] Classical test theory equating analyses or item response theory calibration and linking procedures may be used to build and maintain the score scale over time. These topics are beyond the scope of this discussion.
[5] The *No Child Left Behind Act* of 2001 mandated that state examinations have a minimum of three categories.

goals and rigor for the final performance standards; (b) Content PLDs that are more detailed and ideally aligned to content claims; (c) Range PLDs that are grade-, cognitive- and content-specific descriptors that reflect requisite knowledge, skills, and processes that can guide item writing for particular proficiency levels; and (d) Threshold PLDs that are used to guide standard setting (i.e., specifying minimum performance-level expectations). This framework provides a reasonable way to think about the use cases for PLDs and the notion of Range PLDs helps to somewhat tie performance expectations to item writing and test assembly. However, there is still somewhat of a disconnect between item and test design, item writing, test assembly, psychometrics—including equating, scaling, and scoring practices—and score interpretation. It also remains unclear the degree to which PLDs are legitimately useful for score interpretation, despite the logical connections made explicit via standard setting to set cut scores based on the PLDs.

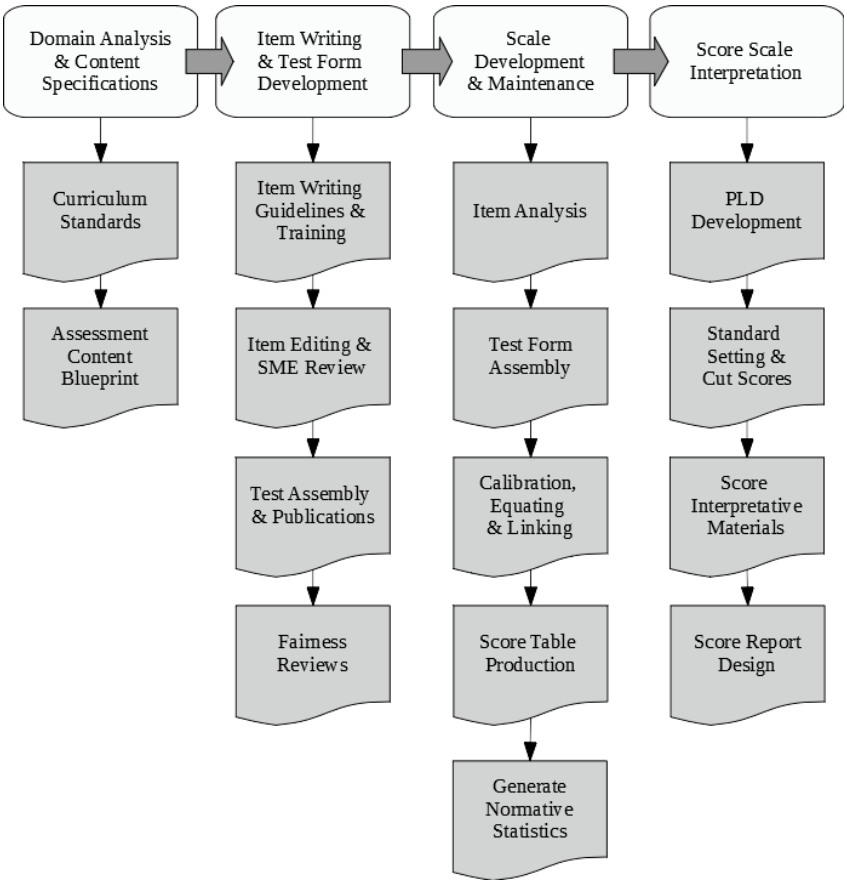
Given the lack of clear alignment between PLDs and item/test design and development, we might attempt to instead distance the interpretation of test scores from the PLDs per se, as the SBAC has suggested on their website:

> Although Achievement Level Descriptors are intended to aid interpretation of the four categories, they are less precise than scaled scores for describing student growth or changes in achievement gaps among groups of students since they do not reveal changes of student scores within the four levels. Thus, achievement levels should be understood as representing approximations of levels at which students demonstrate mastery of a set of concepts and skills. (SBAC, n.d.)

However, this suggestion and the rather unusual comment about the classification levels being "less precise" raises potentially serious questions about the PLDs having any practical utility insofar as interpreting student performance on a test to the expectations reflected by the PLDs.

The fundamental dilemma seems to stem from the way that we design and implement operational testing programs and ultimately develop our score scales. Most test forms and the score scales are usually constructed by "siloed" procedures carried out by different groups within an organization doing different things at different times without extensive coordination or careful integration when it comes to validity and score use. Test developers tend not to worry about the statistical properties of the score scale, psychometricians often view test content issues as an annoyance that gets in the way of optimizing certain statistical score scale properties, and end users must then rely on PLDs and standard setting and *ad hoc* score reporting techniques to build some semblance of meaning into the score scales. This dilemma is exemplified by Exhibit 2 which shows four core assessment "systems": (i) a system for domain analysis and developing the test content blueprints; (ii) a system for item writing and test development; (iii) a system for score scale development and maintenance (i.e., psychometrics); and (iv) a system for score scale interpretation (PLD development and standard setting to score reporting). The research within each system can be highly sophisticated and thorough. However, in practice, there is typically very little integration among these four systems. Outputs from one system may become the inputs to the next system.

Exhibit 2. Four Core Systems for Operational Testing



The score scale needs to be the focal point of test design! We all-too-often rely on sophisticated statistical or psychometric procedures to devise a set of ordered numbers—the score scale—and then spend enormous time, effort, and money attempting to add meaning to that score scale. From a principled design perspective, if we want meaningful interpretations, we need to *design and build the score scale to support the desired statistical properties and the intended interpretive arguments*!

Section 4.0: Systems Integration: Toward an Understanding of the NGSS Score Scales

The NGSS are detailed performance standards for what K-12 students should know and be able to do. They are based on the *A Framework for K-12 Science Education* developed by the National Research Council (NRC, 2012). The standards are grade-specific through fifth grade and organized into grade bands for the higher grades. The NGSS were developed by a collaborative effort between 26 states and numerous partners across the US in an effort to improve science and engineering education.

Various states have started to implement the NGSS into their curricula and assessment blueprints. However, the NGSS represent a *table rasa* where there is a unique opportunity to break out of the siloed, four-system model of operational testing described in Section 3.0 and move toward a more integrated system of test design, task production, psychometrics, and score interpretation. PLD development needs to be a part of that integrated system.

Section 4.1: What are the NGSS?

As noted above, the NGSS are performance expectations based on the premise that science instruction should include an intersection of practice, content, and connection. It is essential to realize that in the same way that the NGSS are not proposed as curriculum, neither are they a set of PLDs nor test specifications. The NGSS are detailed K-12 performance standards that reflect three distinct and equally important dimensions to learning science per the NGSS: (1) DCIs; (2) SEPs; and (3) CCCs. These three dimensions have been combined to form each standard—or performance expectation—to promote classroom learning experiences that help students build a cohesive understanding of science over time.

> **DCIs**—sometimes called the "big ideas"—are the key ideas in science that have broad importance within or across multiple science or engineering disciplines. These core ideas build or scaffold upon each other as a student progresses through grade levels. They are grouped into the four domains of physical sciences, life sciences, Earth and space sciences, and engineering and technology.
>
> **SEPs** describe what scientists do to investigate the natural world and what engineers do to design and build systems (e.g., designing controlled experiments). The practices better explain and extend what is meant by "inquiry" in science and the range of cognitive, social, and physical practices that it requires. Students must engage in these practices to build, deepen, and apply their knowledge of core ideas and CCCs.
>
> **CCCs** help students explore connections across the four domains of science, including physical sciences, life sciences, Earth and space sciences, and engineering and technology (e.g., "cause and effect"). The CCCs help students develop a coherent and scientifically-based view of the world around them.

Exhibit 3 attempts to capture the complexity of the NGSS framework. The NGSS performance expectations addressing the four domains are multi-dimensional and represent varying combinations and intersections of the DCIs, CCCs, and SEPs. The notion is that we cannot consider the DCIs without also considering the CCCs and SEPs.

Exhibit 3. A Graphical Representation of the NGSS Framework

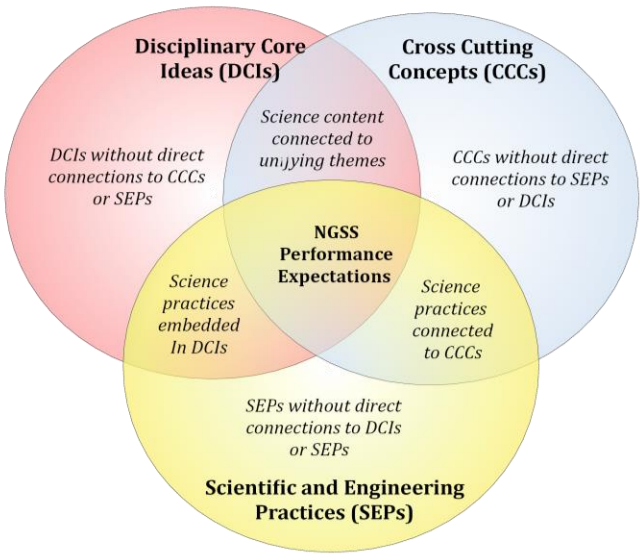
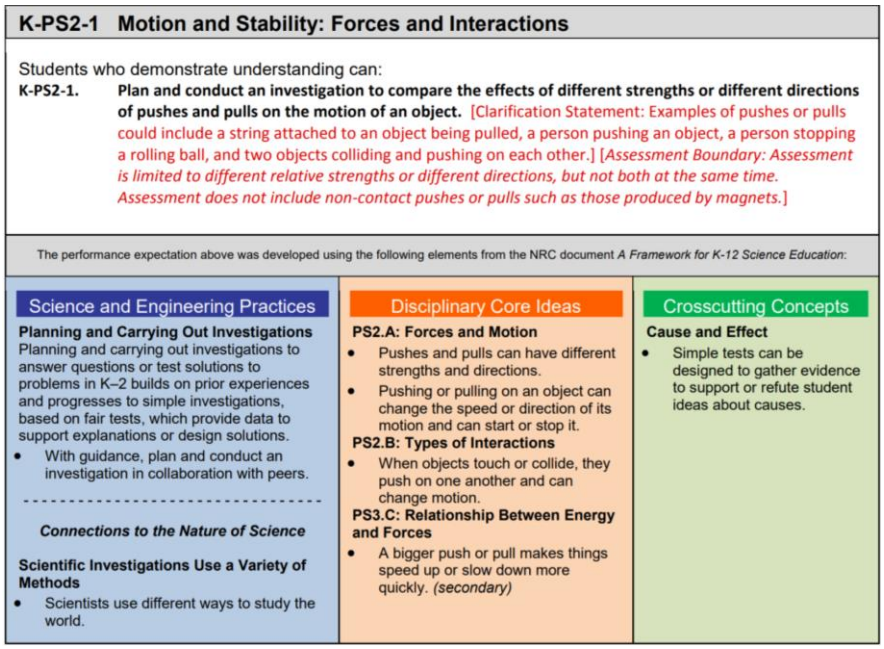


Exhibit 4 shows a single NGSS performance expectation for Physical Science, *Motion and Stability*. The performance expectation is shown at the top of the chart (K-PS2-1). The three NGSS dimensions embodied in the performance expectation are shown in the middle section of the chart: SEPs in blue; DCIs (content) in orange; and CCCs in green. The bullets in this section communicate the elements of the dimensions. The bottom of the chart shows connections to other DCIs and to the Common Core State Standards.

Exhibit 4. A Sample NGSS Performance Expectation for Physical Science[6]



K-PS2-1 Motion and Stability: Forces and Interactions

Students who demonstrate understanding can:

K-PS2-1. **Plan and conduct an investigation to compare the effects of different strengths or different directions of pushes and pulls on the motion of an object.** [Clarification Statement: Examples of pushes or pulls could include a string attached to an object being pulled, a person pushing an object, a person stopping a rolling ball, and two objects colliding and pushing on each other.] [*Assessment Boundary: Assessment is limited to different relative strengths or different directions, but not both at the same time. Assessment does not include non-contact pushes or pulls such as those produced by magnets.*]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices

Planning and Carrying Out Investigations
Planning and carrying out investigations to answer questions or test solutions to problems in K–2 builds on prior experiences and progresses to simple investigations, based on fair tests, which provide data to support explanations or design solutions.

- With guidance, plan and conduct an investigation in collaboration with peers.

Connections to the Nature of Science

Scientific Investigations Use a Variety of Methods

- Scientists use different ways to study the world.

Disciplinary Core Ideas

PS2.A: Forces and Motion

- Pushes and pulls can have different strengths and directions.
- Pushing or pulling on an object can change the speed or direction of its motion and can start or stop it.

PS2.B: Types of Interactions

- When objects touch or collide, they push on one another and can change motion.

PS3.C: Relationship Between Energy and Forces

- A bigger push or pull makes things speed up or slow down more quickly. *(secondary)*

Crosscutting Concepts

Cause and Effect

- Simple tests can be designed to gather evidence to support or refute student ideas about causes.

If we were to adopt a traditional test design and development process, we might bundle the content from the performance expectation shown in Exhibit 4 with other NGSS performance expectations from the same grade or grade band, prioritize the content across the performance expectations, and then develop and refine a content outline that would ultimately become the content blueprint. That is, we would write and then field test enough items to build several test forms. Items with reasonable statistics—especially high item-total score correlations would probably be retained for operational use. We would next construct and administer the test forms—ideally matching our established content blueprint as well as statistical form-assembly targets such as hitting an average item difficulty target, meeting a minimum score scale reliability coefficient, or satisfying an item response theory test information function target. Finally, we would develop the score scale and might then set our NGSS-based cut scores on the score scale for three or more levels—including clarifying the requisite PLDs.

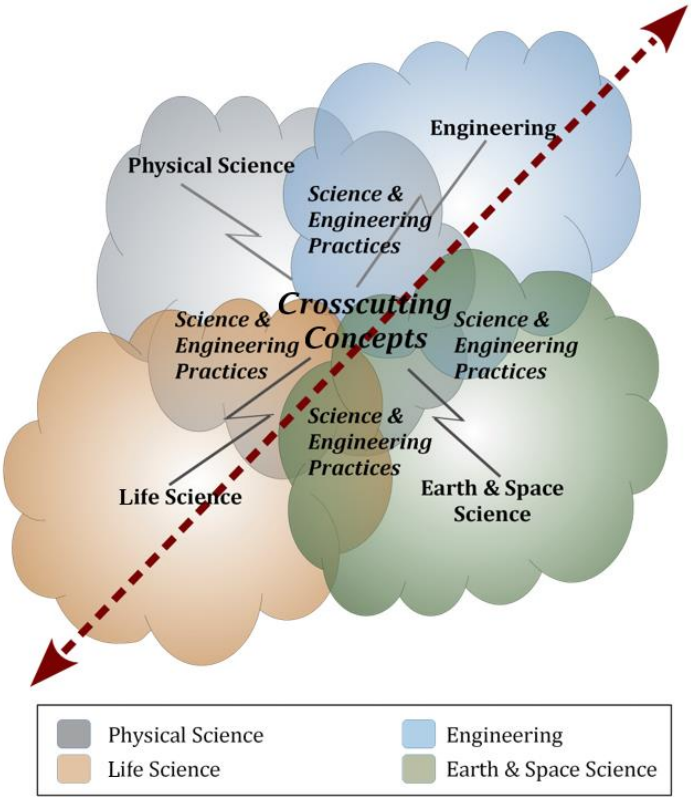
However, that traditional approach would not necessarily produce overly useful scores beyond the usual array of normative interpretations and policy-based decisions usually tied to the proportions of students classified into performance levels. The fundamental problem is that the interpretation of the score scale is NOT viewed as a priority from the onset—before a single test item is created. If it were, we would *focus the majority of our efforts on ensuring that the scores—point-by-point—provided tangible and*

[6] Sample from: NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

concrete performance-based evidence about what students know and are able to do within the context of our assessments!

Exhibit 5 shows an apparent complication in the context of the NGSS. The NGSS, while fairly well documented for purposes of guiding instructional design and curricula, are NOT assessment *progressions per se.* The cloud graphics in Exhibit 5 represent the four domains—undifferentiated by grade or grade band (see the legend in the exhibit). CCCs and SEPs are also shown. Combined, these central components of the NGSS represent an elaborate "learning space" in which students and teachers need to operate. However, from an assessment perspective, consider that we are only allowed to project a single line or "vector" through this space[7]—the score scale. That is the dashed double-headed arrow that cuts through the NGSS learning space.

Exhibit 5. A Score Scale Projection Through the NGSS Learning Space



Since the score scale cannot cover the entire NGSS learning space, what should it represent as a progression of knowledge and skills? The NGSS provide no guidance in that respect. Instead, states and other partners must develop their own trajectories—the make-up of the score scale(s) that cuts through the learning space[8]. However, as noted in Section 1.1 (see Envisioning the Score Scales and Articulating the PLDs for Each Score Scale), the first decision a state must make is to determine the number of score scales they intend to have. Next, a state must carefully articulate the proficiency claims and evidence requirements along the score scale(s) and their relation to the standards. The PLDs can provide useful

[7] Multiple score scales could be devised (e.g., one scale per grade and per DCI).

[8] The phrase "learning progressions" is also specifically avoided here to avoid confusion between documented progressions for teaching and learning versus the progression of task-based challenges that can be included on most standardized tests.

connections between what can be measured along the score scale(s) and the NGSS grade-specific performance expectations. It is simply nonsensical to discuss PLD developments in the absence of that vision.

Section 4.2: Is There a Better Way to Conceptualize PLDs for the NGSS?

A paradigm shift is slowly happening in testing toward principled assessment design (Hendrickson, Huff, & Luecht, 2010; Ferrara, Lai, Reilly, & Nichols, 2017). Spearheaded by the seminal work of Messick (1989, 1994), Mislevy (1996) and Mislevy, Steinberg & Almond (2003), *Evidence-centered Design* (ECD) and similar frameworks have started to reshape the assessment landscape. This paradigm shift can be characterized by three important features. The first is a construct-focused view of assessment design where we develop a clear understanding of the trait being measured within the tested population, the detailed proficiency claims that we want to make about student performance, and the nature of assessment evidence that we would need to see to justify our claims (Mislevy, 1996; Mislevy, et al 2003). The second is a strong task-oriented perspective on score scale design where test items (tasks) are developed using carefully manipulated design features—much like controlling conditions in scientific experiments—to elicit the requisite construct-relevant evidence (Williamson & Bauer, 2013; Luecht, 2012, 2013; 2014). The third is a recognition that almost all score scales have some basic statistical characteristics related to ordering principles. Those same ordering principles need to be reflected in the test content blueprints and our item writing practices. It is not sufficient to rely wholly on sophisticated statistical modeling and large samples of data to extract a set of ordered numbers and only then concern ourselves with what the numbers mean.

One way to take on the complexity of assessment design and its relationship to PLD development for the NGSS is to design what has been called a *Conceptual Assessment Framework* (CAF; Mislevy, Steinberg, & Almond, 2003; Mislevy & Riconscente, 2006; Mislevy & Haertel, 2006; Mislevy, Steinberg, Almond, & Lukas, 2006). The CAF is based on a domain analysis that describes the content knowledge and skills to be measured. The CAF also further articulates the performance claims and performance-based evidence required to support those claims (Mislevy & Haertel, 2006).

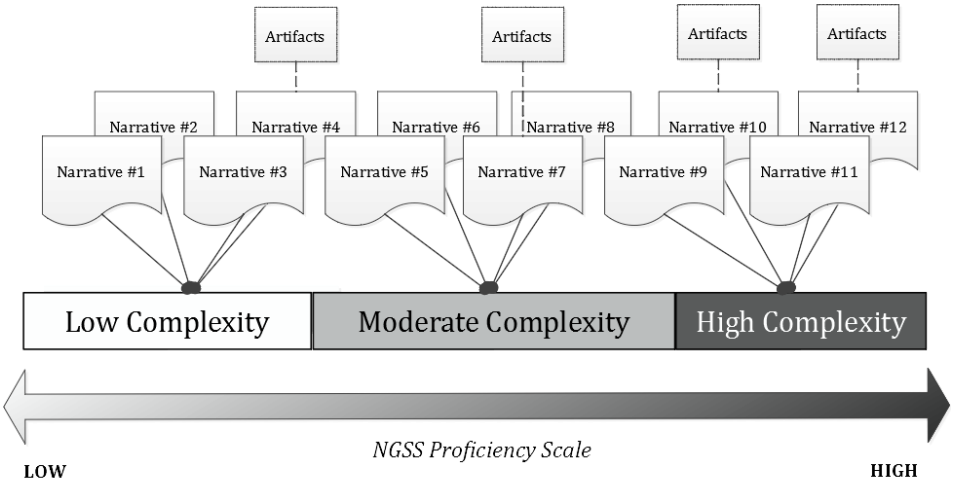
In an attempt to provide concrete direction as to how to get from a CAF to a score scale, Luecht (2006, 2012, 2013, 2014) has framed the problem as one of engineering design where we need to move from a concept to a solution in an efficient and effective way. This Assessment Engineering (AE) framework includes two important concepts: the *construct map* and the *task model map*. A construct map lays out the knowledge and skills, as well as documents the proficiency claims and evidence required to warrant the claims. Wilson (2005) calls this process “construct mapping”. As he notes in describing a construct map,

> It’s most important features are that there is (a) a coherent and substantive definition of the content for the construct; and (b) and idea that the construct is composed of an underlying continuum—this can be manifested in two ways—an ordering of the respondents and/or and ordering of item responses. (p. 26)

A construct map is therefore an elaborated design specification for the ordered knowledge and skill claims that we wish to make about some proficiency of interest. The required evidence to support each claim should likewise be included in the construct map as exemplified in Exhibit 6. Narratives and artifacts can provide concrete examples of the type of measurable evidence needed to support a specific proficiency claim. Defining a skill and knowledge domain such as “eighth grade science” is not a construct definition because it does not specify any claims in terms of cognitive skills and knowledge

structures, nor does it imply any type of ordering of skills and content knowledge along some score scale. Neither is a typical test content blueprint a complete design specification because there is usually no ordering of the specified content by difficulty nor complexity level.

Exhibit 6. Narratives and Artifacts as Part of the Construct Mapping Process



A task model map lays out the discrete distribution of test content (i.e., task descriptions that are designed to be different in complexity and simultaneously directly related to the proficiency claims that define the construct map). The task model map effectively replaces the traditional content blueprint with complexity-ordered, content-anchored task specifications that can guide item writing. This task model map presents a well-articulated progression of tasks that elicit the requisite knowledge and skills needed to justify the proficiency claims.

When test forms are constructed using this type of principled-design approach, the meaning of "progress" along the score scale will mirror the layered progressions along the construct map. Following the task model map in terms of both item writing and test form assembly further ensures that statistical score scale properties such as precision estimates[9] are embodied in the design. In short, this approach to test design ensures that: (a) the required performance evidence from the scored item responses supports the proficiency claims; (b) every item is created following a detailed design specification that represents an appropriate level of the NGSS content knowledge and skills specific to a particular "location" on the score scale; and (c) the statistical properties and interpretation of the scale score points are fully consistent with the measurement intent.

So, where do PLDs fit into this picture? PLDs can be "attached" to the construct map to denote policy-mandated expectations about the levels of performance. More importantly, they can be carefully articulated relative to the complexity of content-based descriptors making up the task model map so that they will have a direct and strong relationship to the score scale. Standard setting activities can still take place, but the result will be less ambiguity between the language of the policy PLDs, the test

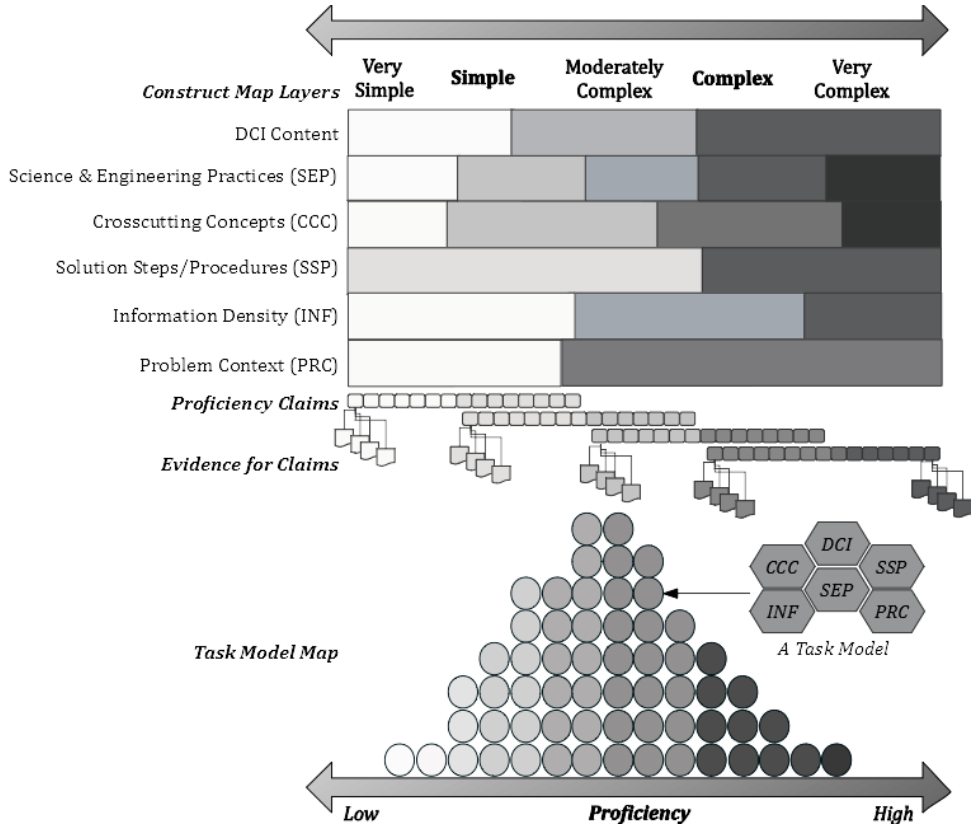
[9] The concentration of "measurement information" along a score scale is directly related to the precision of the estimated scores—often quantified as "conditional standard errors." We would increase the number of task models at a particular score scale location—for example, at the cut scores separating adjacent proficiency categories—to reduce the conditional standard errors in that region of the score scale.

content, and the properties of the score scale. Furthermore, referring to the four categories of PLDs suggested by Egan, Schneider, & Ferrara (2010), the construct map and task model map would include sufficient detail to represent Content, Range, and Threshold PLDs. Everything is done *by design!*

Exhibit 7 shows the nature of the synergy between a construct map and the corresponding task model map. The construct map at the top of the exhibit is composed of *layers* that each denote a content or cognitive complexity dimension. For example, there is one layer dedicated to DCI content (e.g., physical science, Earth and space science, life science, or engineering and technology). The content would be ordered from simpler concepts to more complex networks of core content concepts. Similarly, there are layers for SEPs and CCCs. Finally, there are three layers that describe the implied level of problem-solving given to the number and complexity of procedures or steps to a solution (SSP), the varied amount (density) of information (INF), and the complexity of the setting or problem in which tasks are to be performed (PRC). Each layer has two or more levels going left to right from relatively simple levels to more complex levels. The layers, in turn, are connected to proficiency claims (referring to the ▭ symbols in Exhibit 7) and supporting evidence requirements (see ▭ symbols—also see Exhibit 6).

The construct map layers, proficiency claims, and evidence requirements are used to create task models. The task model map in this example (bottom of Exhibit 7) has 60 task models distributed across the score scale. The highest concentration of tasks occurs near the center implying a need for increased measurement precision in that region of the score scale. The distributional shape of the task model map is a design choice. The number of task models corresponds roughly to the number of test items and we can allocate the concentration(s) or density as needed.

Exhibit 7. Connections Between a Construct Map and Task Model Map



Conceptually, it is straightforward—we specify the task model map that contains detailed task descriptions for each task model, and then use the task models to generate test items that reflect the intended evidence-based design. Each task model is an in-depth content and cognitive specification for designing items that meet the evidence requirements at that point along the construct map. For example, a task model "molecule" (see ⬡ shapes in the lower section of Exhibit 7) is shown for one task model. The task model specification is a composite of the complexity that would be indicated by the choices of DCI, SEP, CCC, SSP, INF, and PRC features drawn from the six layers above. This particular task model would be moderately difficult because it appears in the upper part of the task model map. The task models differ in their complexity (easier to harder tasks) as we move left to right, following the intended progression of claims along the construct map, above.

This type of principled design approach sets up the PLD development to be fully aligned with the content of each test and the properties and interpretation of the score scale points. Consistent with recommendations by Huff and Plake (2010), we want our PLDs to be: (i) sufficiently detailed (granular) in size to be useful; (ii) well understood by the end users (teachers and educators, at least); (iii) exacting in their description of the skills and knowledge required across the levels; and (iv) supported by clear evidence requirements related to the actual assessment content. In short, this type of design plan ensures that the PLDs will provide useful performance expectations relative to test content and the meaning of the scores. Trying to articulate PLDs without this type of principled design—including the extensive detail about the content-based meaning of the score scale—is like addressing a heart-felt love letter, "*To whom it may concern*."

Section 5.0: Moving Forward

Useful and effective PLDs cannot be developed in a policy-focused vacuum. There needs to be clear connections between the PLDs and the accountability, instructional, or other inferences made from them. There also needs to be unambiguous links between the PLDs and the domain standards like the NGSS, the test content blueprints, the items and, ultimately the cut points on a score scale.

The challenges are fairly well understood. We need to: (a) choose the optimal specificity and grain size for the PLDs (i.e., navigate through the "mile-wide and inch-deep" dilemma); (b) document their usefulness and salience (ideally for learning and instruction); (c) keep in mind what is easily observable or measurable as evidence on a typical assessment; (d) make appropriate assumptions about opportunity to learn, prior learning, and prior knowledge; and (e) recognize the inherent assessment needs to develop and document a clear progression of incrementally more advanced and complex test content to support proficiency claims. Ultimately, we need to accept that merely layering *post hoc* meaning or inference onto a score scale is a poor substitute for good construct design.

The amount of effort and resources needed to implement a principled assessment design cannot be trivialized. It is complex and requires a great deal of careful, iterative work. Of course, the alternative is to follow a more traditional approach. That second strategy would produce a foregone outcome perhaps best summarized by a statement by Albert Einstein when he reportedly said, "If you always do what you always did, you will always get what you always got[10]."

[10] Goodreads.com. (2018). *Albert Einstein Quotes (1114 quotes)*. [online] Available at: https://www.goodreads.com/quotes/1171726-if-you-always-do-what-you-always-did-you-will.

There are two approaches that can be used to implement a principled assessment design approach and associated PLDs. One approach would involve developing the assessments using ECD or AE where expectations for performance are set later in the process in conjunction with standard setting (Hambleton, 2001; Cizek & Bunch, 2007; Perie, 2008; Egan, Schneider, & Ferrara, 2012). The primary difference from traditional standard setting and PLD development would be that the score scales and incremental progressions of task complexity along the construct maps could provide greater contextual detail for interpreting the PLDs and standards. However, a caveat with this approach is that the task rigor and expectations for performance would merely be "attached" to the score scale in somewhat of a *post hoc* way. The ideal approach would be to integrate the PLD development into the construct and task-model mapping activities. Under this approach, the PLDs would be functionally embedded within the score scale. This approach has the additional benefit of helping to guide the distribution of the task models since we would like the highest concentration of task models on the map to fall near the cut scores as a means of maximizing the accuracy of the proficiency classifications.

Regardless of the approach used, an actionable principled assessment framework for PLD development requires an upfront commitment to iterative design principles. Much like any manufacturing engineering design process (e.g., Dym & Little, 2009; Dieter & Schmidt, 2012), we can conceptualize six phases for principled assessment design: (i) research and initial documentation; (ii) articulating the construct and score scale design requirements; (iii) evaluating practical constraints, assumptions, and feasibility[11]; (iv) conceptualization and prototyping (construct mapping and task model mapping); (v) preliminary design; (vi) detailed design; (v) implementation with alpha and beta testing for refinement of the assessment production tools and resources; and (vi) future design version planning and communication.

Unlike a traditional standard setting meeting (Hambleton, 2001), these types of integrated activities cannot be carried out during a conference of educational policy makers over several days. Rather, they require dedicated teams of NGSS subject-matter experts, educators, cognitive scientists, psychometricians, and policy makers going through the iterative design process over a reasonable period of time with the end goal of an interpretable and psychometrically sound score scale with defensible PLD-based cut scores. Furthermore, with any design process, it is important to adopt a versioning strategy and a commitment to finish "version 1.0." A balance is needed. On one hand, we want the process to be careful and thorough. However, if too much time is spent researching or developing initial design specifications, the energy and resources needed for subsequent phases may wane. The commitment to iterative design ensures that any design deficits identified later in the process can be detected and appropriate design modifications implemented.

There are five key questions to address in moving forward with the SCILLSS project for the NGSS. First, who should be involved in the development process? As noted above, development teams are needed—at least by grade or for grade bands—and should include NGSS subject-matter experts, educators, cognitive scientists, psychometricians, and policy makers. Second, what materials will they need? It is not sufficient to merely hand-off the NGSS to the designated groups and say, "Create a score scale and PLDs." The same development teams need to be involved in the initial research all the way through at least the detailed design phase of the process. Third, what will they do (actionable procedures)? It is likely that the design teams will need to go through training on the ECD or AE design process. Procedures, protocols, and communication channels need to be developed to allow the development

[11] Feasibility is an essential criterion to ensure that the performance expectations and evidence can be represented by typically test items—whether selected- or constructed-response tasks. See Huff, Steinberg, & Matts (2010) for a discussion of this issue relative to the College Board's Advanced Placement™ examinations.

teams to function remotely if needed. As noted with respect to standard setting, this type of activity cannot be handled in a brief meeting of policy makers or educators. Fourth, how will the process be documented and evaluated? When implemented as a type of engineering design process, the documentation becomes an integral part of the process. Small scale empirical studies can also help ensure that any preliminary score scale and PLD designs are functioning as intended. Finally, what will be the final outcomes and how will those outcomes be used moving forward? Ideally, this type of process will generate NGSS-anchored test score scales with clear connections between the intended construct definitions and proficiency claims, the evidence requirements, the complexity-ordered, measurable test content, the item (task) designs, the properties of the score scales, and the policy-based performance expectations.

Implementing a principled-design approach such as ECD or AE may seem like a daunting endeavor, and it is worth acknowledging the hard work and resources required for such an approach in comparison to traditional test design. However, this up-front investment in the design process will establish resources and processes that will make future development work less burdensome, and will result in assessments that produce meaningful, interpretable scores that are useful to stakeholders.

Section 6.0: References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Bejar, I., Braun, H. I, & Tannenbaum, R. J. (2007). A prospective, predictive, and progressive approach to standard setting. In Lissitz, R. (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting*. Maple Grove, MN: JAM Press.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Dieter, G. E., & Schmidt, L. C. (2012). *Engineering design, 5th edition*. New York: McGraw-Hill Education.

Dym, C.L., & Little, P. (2009). *Engineering design, 3rd edition*. New York: John Wiley & Sons, Inc.

Egan, K. L., Scheinder, C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In C. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 79-106). New York, NY: Routledge.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2017). Principled approaches to assessment design, development, and implementation. In A. A. Rupp & J. P. Leighton (Eds.), *Cognition and assessment* (pp. 41-74). New York, NY: Wiley Blackwell.

Goodreads.com. (2018). *Albert Einstein Quotes (1114 quotes)*. [online] Available at: https://www.goodreads.com/quotes/1171726-if-you-always-do-what-you-always-did-you-will

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspective*s (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement, 4th edition* (pp. 433–470). Washington, DC: American Council on Education.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K.W. (2016). *Constructing assessment tasks that blend disciplinary core Ideas, crosscutting concepts, and science practices for classroom formative applications.* Menlo Park, CA: SRI International.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*(4), 358-377.

Huff, K., & Plake, B. S. (2010). Innovations in setting performance standards for K-12 test- based accountability. *Measurement: Interdisciplinary Research & Perspective, 8*, 130-144.

Huff, K., Steinberg, L. S., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*, 310-324.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement, 4th edition* (pp. 17-64). Washington, DC: American Council on Education.

Linn, R. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Educational Policy Analysis Archives*, *13*(33). Retrieved 13 October 2018 from https://epaa-asu-edu.libproxy.uncg.edu/ojs/article/view/138/264

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspective*s (pp.175-217). Mahwah, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2006, May). *Engineering the test: Principled item design to automated test assembly*. Invited special event presentation at the Annual Meeting of the Society for Industrial and Organizational Psychology.

Luecht, R. M. (2012). An Introduction to Assessment Engineering for Automatic Item Generation. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation* (pp. 59-101). New York: Taylor-Francis/Routledge.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14*.

Luecht, R. M. (2014). Computerized adaptive multistage design considerations and operational issues. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: theory and applications* (pp. 69-83). Boca Raton, FL: CRC Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, *3rd edition* (pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *32*(2), 13-23.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379-416.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues & Practice, 25*, 6-20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S.M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-66.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, NJ: Lawrence Erlbaum Associated.

National Center for Educational Statistics. (2012). *The NAEP Science Achievement Levels*. Retrieved October 12, 2018 from https://nces.ed.gov/nationsreportcard/science/achieve.aspx#2009_grade8.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Perie. M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*, 15-29.

Plake, B., Huff, K., & Reshetar, R. (2010). Evidence centered design as a foundation for ALD development. *Applied Measurement in Education, 23*, 307-309. Retrieved from

Smarter Balanced Assessment Consortium. (n.d.). Retrieved October 20, 2018 from the SBAC website: http://www.smarterbalanced.org/assessments/scores/.

Williamson, D., & Bauer, M. (April, 2013). *An evidence centered design for assessment of technology and engineering literacy*. Paper presented at the Annual NCME Meeting, San Francisco, CA.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of naep score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359-375.

Zieky, M., Perie, M., & Livingston, S. (2008). *Cut scores: A manual for setting performance standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.