Considering Systems of Assessment in the Context of Three Dimensional Science Standards: A Focus on District and State Layers

September, 2018

Nathan Dadey
The National Center for the Improvement of Educational Assessment

A Paper Commissioned by the Technical Issues in Large Scale Assessment (TILSA)
State Collaborative on Assessment and Student Standards (SCASS)
of the Council of Chief State School Officers (CCSSO)

Suggested Citation:

Dadey, N. (2018, September). *Considering Systems of Assessment in the Context of Three Dimensional Science Standards: A Focus on District and State Layers.* Washington, DC: Council of Chief State School Officers.

Executive Summary

Overall Summary

The 2014 National Research Council (NRC) report, *Developing Assessments for the Next Generation Science Standards,* presents vision of a system of assessments to support teaching and learning. This paper attempts to operationalize at least a part of that vision by articulating some the key design decisions involved in the development of the state and district levels, or layers, of a system of assessments. These key design decisions deal with multiple areas – the design of the assessments within each level, the frameworks that defines the content of the assessments, the way in which student performance is modeled and the ways in assessment results are used – that are relevant to systems design in any domain (e.g., science, English Language Arts, mathematics). However, the decisions addressed are by no means exhaustive. Instead, the design decisions are targeted towards some areas that are particularly challenging to address in light of the three dimensional nature of the NGSS.

By considering these design decisions, those working on systems of assessments may be able to better characterize the set of assessments currently in place and then identify which, if any, of these areas of design should be singled out for improvement. In particular, considering the design decisions made at the state and district layers in relation to one another may suggest improvements that allow each layer to better complement the other. Given that the state layer is almost always less flexible than the district layer, a pragmatic approach is to design the district layer to complement the state layer. Finally, this brief does not address many of the other challenges associated with the implementation of a system of assessments, like issues related monetary constraints, communications, political boundaries, operational psychometrics.

Overview

The narrative of this paper is dedicated to exploring, in detail, a number of design decisions involved with implementing assessments within the state and district layers, with the aim that such detail can aid considerations across the two layers. However, the level of detail provided may obscure the important concepts within the work. The overview below presents the concepts important to this work in a streamlined format to both provide a succinct description of the ideas and to clarify the organization of the paper.

Detailed Summary

I. Systems of Assessment

- The idea of a “system of assessments” has been gaining in popularity, especially since the publication of *Knowing What Students Know* (Pellegrino, Chudowsky & Glaser, 2001).

    - In just five pages, Pellegrino et al. (2001) outline a plan and criteria[1] for "*coordinated systems* of multiple assessments that work together, along with curriculum and instruction, to promote learning" (p. 252, original emphasis), calling systems that meet these criteria "balanced." In addition, although not explicitly mentioned, the authors imply that such a balanced system ideally spans districts, schools and classrooms within a state.
    - The idea of a system of assessments has been addressed in various works since Pellegrino et al. (2001), with each author providing a slightly different definition of what a system of assessment is (e.g., "comprehensive" Perie, Marion, & Gong; 2009; Ryan, 2010; or "next generation," Darling-Hammond & Pecheone; 2010; Herman, 2010). The 2014 National Research Council (NRC) report, *Developing Assessments for the Next Generation Science Standards*, stresses that not only is a system of assessments desirable, it is a *necessity* given the complexity of the Next Generation Science Standards.
- Despite the interest and work on systems of assessment, there are few, if any, examples of systems (Conley, 2018) that meet the vision laid out by Pellegrino et al. (2001) and expanded on in later work (e.g., NRC, 2014).
    - This is not to say that development on systems of assessment has not been taking place, but rather that systems of assessments that span classroom to state layers, or even a subset there within, is extremely challenging.
- This work attempts to operationalize at least a part of the vision laid out by the 2014 NRC report, by articulating some the key design decisions involved in the development of the state and district levels, or layers, of a system of assessments.
    - These design decisions are meant to allow those considering systems of assessments to characterize the set of assessments currently in place and then identify which, if any, of these areas of design should be targeted for improvement. By examining the state and district layers together, improvements can be implemented with the aim of increasing the complementarity of each layer with the other.
- Before addressing the state and district layers in detail, this work examines the idea of a system of assessments and:
    - Conceptualizes a system of assessments as being made up of multiple levels or layers, with an emphasis on the idea that there is no *one* system of assessment within a state.
    - Suggests that models of student learning, upon which a system is supposed to be based, are extremely difficult to coordinate across levels, particularly the state and district levels. Models of student learning tend to vary, to different degrees, across classrooms, schools and districts within a state.

[1] These criteria are comprehensiveness, coherence and continuity (Pellegrino et al., 2001). Comprehensiveness means that multiple sources of information collected through different processes (e.g., standardized assessments, performance assessments) are considered. Coherence means that the model or models of student learning underlying all of the assessments are compatible and that these same model(s) underlie curriculum, instruction, and assessment. Finally, continuity means that assessments should measure student progress in relation to a model of student progression.

    - Proposes that the uses of assessments within each layer be articulated in detail so that the connections, or lack thereof, between layers can be clearly seen.
- Within both the state and district layer sections, this work considers some key design decisions made in designing a given layer of a system of assessments. These decisions are presented in four categories[2]: (1) assessment design, (2) assessment framework, (3) dimensionality, and (4) use cases.
    - The state layer section focuses on laying out different decisions that can be made during the assessment development process. This approach is meant to provide clarification around the inferences being made about students, as claims developed for state-level assessments are often not specific enough to inform design decisions at the other levels[3]. However, this work does not devote much attention to refining claims, achievement level descriptors, or reporting categories – instead focusing on the design of the state and district layers and the assessments within each layer.
    - The goal is to make clear these design decisions, so that the district layer can be considered in relation to, and hopefully made complementary or coherent to, the state layer. Such complementarity, however, does not mean that the limitations of the state layer be adopted into the district layer – rather the district layer could be created to address many of these limitations.

II. The State Layer

- The state-layer should be well defined, at least relative to the district layer of an assessment system, despite the added complexities of three dimensional science standards. Decisions about the state layer include those related to:
    - **Assessment Design**, specifically whether:
        - The state-level assessments are developed and administered as one time end-of-year or end-of-course assessments, or are more distributed (e.g., through a modular design). No examples were found of the latter approach being implemented by any state.
        - A single assessment covers standards from a single grade or course, or the standards across a grade band.
        - A single assessment is administered once per grade, or once per grade band.
    - **Assessment Framework**[4], regarding whether:
        - The standards are defined as the performance expectations, as the performance expectations plus some additional state defined expectations, or as all possible intersections of the three dimensions of the NGSSS, which are the disciplinary

[2] The terms used for these categories are mostly invented for this paper, but do help organize the decisions into meaningful groups.

[3] Pellegrino et al. call for the "conceptual base for the large-scale assessment [to] be a broader version of the one that makes sense at the finer-grained level" (2001, p. 255). The claims often given for state-level assessments are not clear enough to determine whether basis for the large-scale assessment is a broader or coarse version of that used at lower levels, nor is it entirely clear how to evaluate whether the conceptual basis is the same across levels.

[4] Some might call this the domain definition (see Gong & Norris, 2018), or construct definition.

core ideas, science and engineering practices, and the cross cutting concepts. In addition, the performance expectations alone may be treated as the standards, or the performance expectations and each dimension that defines each expectation (i.e., the foundational boxes).

    - Any aspect of the standards is "foregrounded," or emphasized, to provide additional structure to the claims, achievement level descriptors, score reports, blueprints, item clusters, or any combination thereof.
    - A fixed subset of the performance expectations, a subset of matrix sampled performance expectations across years, or all performance expectations matrix sampled with year are assessed[5].
    - Item clusters are used, and if so, whether the clusters are based on a single performance expectation (one performance expectation is assessed by one cluster) or based on performance expectation bundles (two or more performance expectation are assessed by one cluster). In addition, within performance expectation bundles, the dimensions might be "mixed and matched." This would mean, for example, that a disciplinary core idea from one performance expectation might be assessed with a scientific and engineering practice from a different performance expectation.
- **Dimensionality**, including whether:
    - A unidimensional item response theory (IRT) model is sufficient to summarize variability in student performance (i.e. through a single score scale).
    - There are additional sources of variability that could be summarized via a multidimensional model (including multidimensional item response theory models and diagnostic classification models). Although there a number of multidimensional models that could be used to summarize performance, one model has been suggested for state-level assessments of the NGSS – the the bifactor model (e.g., Gibbons & Hedeker, 199; Rijmen, Turhan & Jiang, 2018).
        - Rijmen et al. (2018) have used the bifactor[6] model to account for within cluster variability, or "cluster effects." These effects arise when there is substantial amount of variability in student item performance is associated with the item cluster groupings (i.e., within item cluster variability is substantial). This clustering of variability poses problems to unidimensional IRT models, as they assume student performance can be summarized well using a latent trait (i.e., a single scale).
        - Within the bifactor model, student performance is summarized by an overall or general scale (i.e., a general latent variable) as well as a cluster specific scale (i.e., a cluster or secondary variable; one for each

[5] This assumes that the state has decided to use the performance expectations to structure the assessment – an assumption made throughout this work. However, this is not the only assumption that could be made.

[6] The bifactor model and the testlet response model proposed by Wainer, Bradlow & Wang (2007) are virtually identical in this context.

cluster). In usual applications of the bifactor model, the focus is placed on the general scale and the cluster specific scales are ignored, as in Rijmen et al.'s (2018) application.

        - However, some of the unique item cluster variance that is ignored may be construct relevant. Thus it is an open question as to whether the cluster variance should be modeled out using the bifactor model, or if a unidimensional model should be used at the risk of violating an assumption of the IRT model. In terms of the latter, Reise, Moore & Haviland (2010) note that, "multidimensional data can yield interpretable scale scores and be appropriately fitted to unidimensional IRT models" (p. 1).

    - **Use Cases**, as to whether:
        - The results are meant to be used as part of the state's system of school identification for targeted or comprehensive support and/or some other specific use. Some possible other uses are detailed in Table 1, but the contents of the table represent only a limited set of examples. Any use should be specified, in detail, so that the reported assessment information can be evaluated in terms of the given use.
        - Ideally, any use should be explicated in a theory of action that couches the use of assessment results in terms of outcomes and the mechanisms which produce the intended effects of an assessment system. Moreover, the theory of action should specify the action mechanisms by which the effects are supposed to be caused, which in turn dictates the interpretations that an assessment or assessments will need to support.
        - A guiding question in examining use is whether the results from the state-level assessment(s) alone are sufficient to support the intended use. The results of the last twenty years of accountability policy would suggest that state-level information alone is not sufficient and that information from other layers of the system should be considered.

III. The District Layer

- The design decisions detailed above regarding the state layer also apply to the district layer. Instead of focusing on those decisions, this section instead focuses on design decisions that may allow the district layer to complement the state layer.
    - As previously noted, such complementarity, however, does not mean that the limitations of the state layer be adopted into the district layer; rather the district layer could be created to address many of these limitations.
    - In addition, there is no clear distinction between the design decisions involved at the state and district levels. Any of the issues could apply to any layer. However, given typical uses, the issues discussed do fit somewhat naturally at the given layers
- Decisions about the district layer include those regarding:
    - **Assessment Design**, specifically:

  - The degree of modularity of the assessment(s) – i.e., the grain size to which the assessment content has been targeted (from a single PE to the complete set of PEs for a grade or grade band) as well as how flexible the timing of administration is (from completely determined by the user to completely determined by the district or state).
    - *Fixed* designs include those assessments that broadly measure a given content domain, a design common in many interim assessments. Such interim assessments are often deemed to have a “summative” or “mini-summative” design, meaning that their blueprints reflect those of the state-level summative assessment (see Gong, 2010 for more on these types of designs).
    - *Block* designs involve multiple assessments, each of which measures large chunks of the content domain (e.g., Smarter Balanced interim assessment blocks, Delaware’s optional unit based assessments).
    - *Modular* designs also involve multiple assessments, but these measure fine grained parts of the content domain (e.g., performance expectation based item clusters like those provide by Kentucky in a task bank).
  - The flexibility of block and modular designs may help alleviate some of the issues that arise due to variation in curriculum and instruction within and across districts.
- **Assessment Framework**, particularly whether:
  - The district-level assessments will take the *same* approach to measuring the standards as the state-level, in terms of depth and breadth, or a *different* approach (Gong, 2010; Herman, 2017; Perie et al., 2009; Brookhart, 2013). There are viable reasons for both approaches.
    - For example, a district might choose to assess intersections outside of PEs assessed on the state-level assessment(s) to signal the breath of the NGSS, more fully represent the standards and thus attempt to mitigate any narrowing of the curriculum.
    - In a different example, the district-level assessments might be used to measure the PEs “as is” while the state-level assessment focuses on transfer of those dimensions captured by the PEs to novel problems or phenomenon (i.e., depth on a limited number of PEs).
- **Dimensionality**, including:
  - What type of measurement model is necessary or desired.
    - Many uses do not require an IRT based measurement model (e.g., if an assessment is being used within a school by a group of teachers instructing the same course, for the purpose of determining whether the overall scope and sequence of instruction should be shifted).
    - IRT based measurement models generally provide (a) a score (be it a scale score or classification) and (b) facilitate comparability so that

comparisons can be made more readily across time or units. Applying an IRT based measurement model to district level assessments is not necessarily more complicated that applying at the state-level, even if a more modular design is desired (see Dadey, Tao & Keng, 2018 for an example in mathematics). The same concerns that arise at the state level are likely to also arise at the district level, including those about multidimensionality,

- - - - A key concern is what type of score is to be reported and how it relates to the state-level. If a scale score or classification is desired, should it be the *same* as that provided on the state-level assessments, or should it differ (i.e., comparability in scores across level)?
    - An alternative or complement to a measurement model is standardized scoring of open ended tasks, possibly with centralized scoring.
  - **Use Cases**, as to whether:
    - The assessment results are meant to be uses for monitoring or instructional purposes. Within the district layer monitoring uses are often conflated with instructional ones, perhaps even more so than within the state-layer. Frequently district-level assessments used for monitoring purposes are also touted as having relevance to teaching and learning, when, at best, these assessments are distal measures of learning with lose and undefined connections to classroom instruction. This is not to say that the district-level assessments cannot be relevant to teaching and learning, particularly when a design can be tailored to instructional sequences and there are supports to do so. However, care does need to be taken to insure that such uses are actually supported by the assessment results.
      - With the right design, any and all of the uses detailed in Table 1 could be supported by district-level assessment.
      - Any use that is evaluative in nature (i.e., used for monitoring) likely needs to be separated from instructional uses. For example, if a fully modular assessment item cluster bank is meant to be used for monitoring purposes, some tasks could be held securely apart from those used for classroom purposes. An alternative approach, might be to simply limit the types of evaluations to low stakes and provide open access to the assessments.
      - The uses across the state and district-levels should be compared for possible disconnects.

IV. Complementarity

- Pellegrino et al. (2001) outline three criteria for evaluating the quality of an assessment system. These criteria are comprehensiveness, coherence and continuity.
- Any combination of these criteria can be used as a guide to improve the overall quality of an assessment system. These criteria should also be considered in light of the categories used to

characterize each layer of the assessment system (assessment design, assessment framework, dimensionality, and use cases).

  - Not all of the three criteria map onto each of the categories well, but considering the interplay is useful.
  - For example, introducing item clusters that are performance based to improve comprehensiveness could better support a variety of uses, and would likely make up for some of the limitations at the state-level. Developing modular district-level assessments, with recommended administration patterns based on common curriculum, could improve coherence.

I. Introduction & Motivation

The idea of a "system of assessments" is an appealing one. That is, by using the results of multiple assessments together – assessments likely developed for different purposes, that provide different information to different users, but all aligned to a common theory of learning – student learning can be improved, and improved in ways not supported through the piecemeal use of assessments (Pellegrino, Chudowsky & Glaser, 2001). The 2014 National Research Council (NRC) report, *Developing Assessments for the Next Generation Science Standards*, calls for the development of such systems of assessments[7] and also calls for such systems to be developed from the classroom level up. Ideally, such systems are meant to span the multiple levels of the educational system, from the classroom to the state. By doing so, all of the assessments within the system – from classroom formative assessments to state-wide accountability assessments – are meant to work *together*.

Meeting this call has proved difficult. Different users of assessment results often hold very different theories of learning and thus conceptualizations of student proficiency. Perhaps most concretely, this can be seen in the numerous scopes and sequences of instruction, driven by differing curricula, found within any U.S. state. Such diversity points out one reason why systems of assessment have been so difficult to implement, because they generally presume that there is one theory of learning uniting a system, when in reality there are different theories operating at each level of the system. This diversity also suggests that district and state initiatives aimed at supporting a system of assessments must be *flexible* enough to work across these differences but also *specific* enough to support a variety of instructional and evaluative uses.

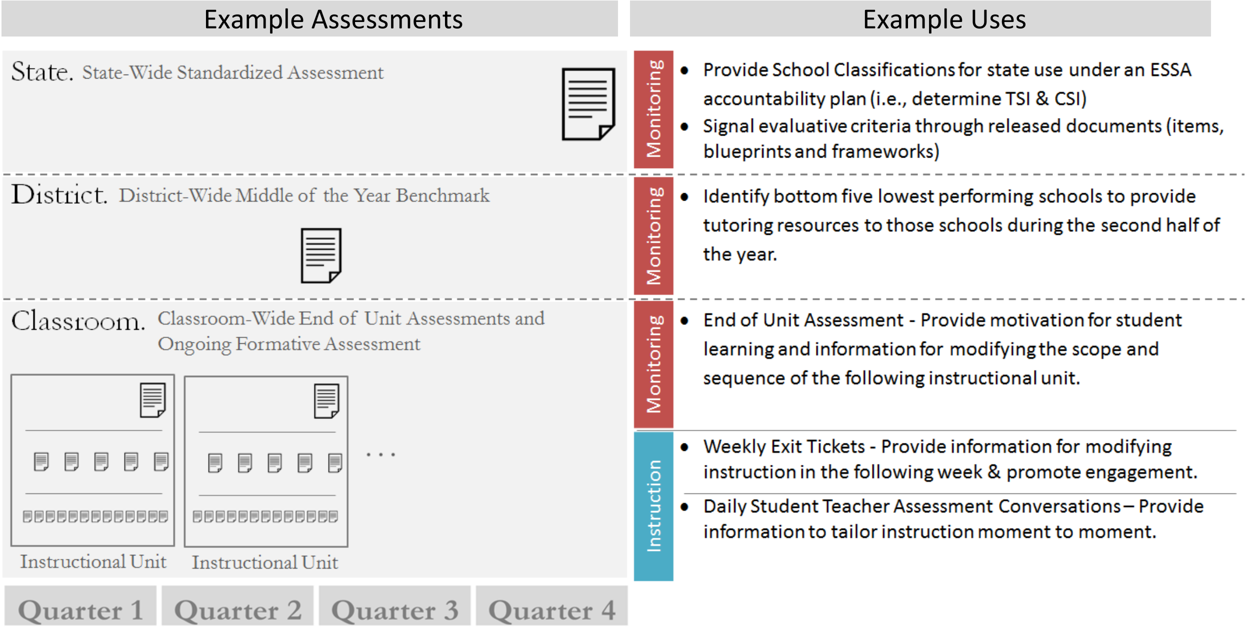
Seeking this flexibility and specificity brings us to the purpose of this work – to examine how the state and district "layers" of an assessment system can be defined and aligned to specific uses, and in doing so hopefully suggest areas in which these layers can be made more complementary to one another. To do so, the next section provides a heuristic that defines a system of assessments in terms of the levels or layers of control within an educational system, as well as how assessment results are used and when. The following two sections then address the state and district-layers in turn. Each section focuses on a limited, non-exhaustive set of design decisions involving the design, framework, dimensionality and uses of the assessments within that layer. These design decisions are relevant to systems design in any domain (e.g., English language arts, mathematics), but the three dimensional nature of the NGSS adds a new level of complexity to these decisions and consequently merits the detailed inspection provided here. In fact, these design decisions are presented precisely because they have been difficult to address in the development of assessments of the NGSS. The final section concludes with some recommendations on ways in which current assessment practices can be move towards the systems approach.

[7] Defined as a coordinated system of "multiple assessments that work together, along with curriculum and instruction, to promote learning" (Pellegrino, Chudowsky and Glaser, 2001, p. 252).

A Guiding Heuristic

This work focuses on the district and state layers of a system based on the idea that articulating the relationships between these two layers could highlights areas for improvement. In doing so, it conceptualizes the levels or layers of an assessment system as the main differentiator between the components of the system. This approach is in line with much of the approach taken by Pellegrino et al. (2001) and subsequent works, like Shepard, Penuel & Pellegrino (2018). However, this focus on levels does differ from the approach taken by the NRC (2014) report, which distinguishes between components of an assessment system by purpose (i.e., classroom assessments, monitoring assessments and opportunity to learn indicators). Perie et al. (2009) also defined the components of a system by their purpose using a somewhat different set of categories (i.e., summative, interim and formative assessments).

Figure 1. *An Illustration of One Hypothetical System of Assessment.*


Example Assessments
Example Uses
State. State-Wide Standardized Assessment
District. District-Wide Middle of the Year Benchmark
Classroom. Classroom-Wide End of Unit Assessments and Ongoing Formative Assessment
Instructional Unit
Instructional Unit
. . .
Monitoring
Monitoring
Monitoring
Instruction
Quarter 1
Quarter 2
Quarter 3
Quarter 4

- Provide School Classifications for state use under an ESSA accountability plan (i.e., determine TSI & CSI)
- Signal evaluative criteria through released documents (items, blueprints and frameworks)

- Identify bottom five lowest performing schools to provide tutoring resources to those schools during the second half of the year.

- End of Unit Assessment - Provide motivation for student learning and information for modifying the scope and sequence of the following instructional unit.

- Weekly Exit Tickets - Provide information for modifying instruction in the following week & promote engagement.
- Daily Student Teacher Assessment Conversations – Provide information to tailor instruction moment to moment.

Notes: A layer for school has been omitted, but figures such as the one above should be used flexibly to illustrate the total set of assessments students take. The classroom level is roughly designed based on Penuel, Frumin, Van Horne and Jacobs's, (2018) example of a set of formative assessments practices coupled to phenomenon based units, in the context of instruction based on the Next Generation Science Standards.

These two different approaches are not mutually exclusive nor should they be. Certain purposes tend to fall at certain layers, e.g., in many cases the district-layer and interim assessments are almost synonymous. Shepard et al. (2018) illustrate this overlap by articulating the common purposes of assessments at each layer of the educational system (see Table 1, p. 26). One way to extend this type of presentation is through the creation of a "map" of the assessments, showing when the assessments at each layer are administered and for what purpose. Figure 1 depicts one *hypothetical* example system of assessment, and also attempts illustrate the system across the academic year. The layers with this map correspond both to how the assessment results are used and also who uses the results. As a hypothetical example, the figure shows the complete set of assessments students take during the year and explains the purposes of each assessment. Such a figure could be improved by incorporating information on curriculum and instruction, illustrating the connections, or lack thereof, between assessments and learning. Such connections would likely take place within the classroom, school and district levels. Levels not shown in the figure could be added, such as school or regional levels.

This type of figure also helps show how a system of assessments can be conceptualized as "configurable." That is, at each layer, decision makers can build, or at least implement, their own set of assessments, designed for particular purposes, resulting in differing systems when taken as a whole. Much of the work on systems of assessments seems to imply there is one monolithic system that spans from the classroom to the state, which could be true if a single school within a single district within a state is considered. Across a state there will be many such systems of assessments.

An example limited to just the district-layer may help. In the same state for a given grade and subject, two different districts may develop different designs. One district may opt to have a single, mandatory assessment administered in the middle of the year to assess some set of knowledge, skills and abilities deemed to reflect learning in the first half of the year and also critical to student performance at the end of the year. Schools with the lowest average scores, say the five lowest performing schools, are then provided with district supported tutors who work inside classrooms and also provide separate drop by tutoring hours during and after school. The other district may provide optional performance tasks, to be administered as needed by educators, with recommendations for ways in which schools or groups of schools can develop communities of practice to use results to inform instruction. These two districts are just one example, within one layer, but illustrate the complexities inherent within a system of assessments.

Looking across all of the layers within a state, many of the systems present, if not most, are likely to not be systems at all, lacking a common theory of learning to unify the system (or what Pellegrino et al. (2001) term vertical coherence). Whether or not a particular set of assessments functions as a system is not the focus of this work, instead the focus is on ways in which state and district assessments can be better designed to complement one another. In one ideal case, district and state assessments would be designed concurrently, with the choices made for one layer feeding back into the other. In most cases, the state-layer assessments are a given and district-layer assessments should be designed or refined to complement the state assessments.

The district layer[8] is particularly important in this work – much has been written on classroom-level and state-level assessment, but less is written on the function of district assessment (although recent work post 2000 has made headway on this, e.g., Abrams, McMillan & Wetzel, 2014; Bulkley, Christman, Goertz, & Lawrence, 2010; Coburn & Talber, 2006; Clune & White, 2008; Davidson & Frohbieter, 2011; Diaz-Dilello, 2011; Shepard, Davidson, & Bowman, 2011; Supovitz & Klein, 2003). To provide clarity on the potential range of uses within the district layer of the system, Table 1 below presents a set of categories generated by Martineau and colleagues (2018), which defines broad categories of use for the information produced by assessments. Note that these categories are not the only ways that various

[8] Although this work is focused on the district-layer, these characterizations apply to any layer between classroom and state (e.g., school, region).

uses can be classified, nor do they capture all of the ways in which assessment results could be used. These categories do, however, provide a starting point. However, these broad categories need to be further articulated to be specific enough to guide practice.

Table 1. *Some Broad Categories of Assessment Use. Adapted from Martineau et al. (2018); used with Permission*.

Category	**Description**
Signal	Maintain a feedback loop between student and teacher to signal next steps Indicate valued knowledge and skills to motivate instruction and student work
Triangulate	Corroborate… formative assessment insights to improve decisions and refine practice unit grades/test results to improve decisions and refine content/scoring marking period grades/test results to improve decisions and refine content/scoring results of a 2+ marking period test to improve decisions and refine content/scoring
Inform Instruction	Monitor… instructional effectiveness for in-the-moment adaptation and course correction student/group needs to differentiate and/or tailor next-lesson planning & instruction student/group needs to differentiate and/or tailor next-unit planning & instruction
Inform Programming	Evaluate achievement to guide… mid- or long-term grouping (including remediation) instructional program placement (e.g., grade, course or track placement)
Grades	Evaluate achievement to support… traditional grading standards-based grading
Eligibility	Evaluate achievement to determine eligibility for… course credit (w/out taking the course) program entrance (e.g., EL, SWD) program exit (e.g., EL, SWD) graduation/diploma annotation formal honors/awards
Readiness	Evaluate… achievement to determine readiness for the next lesson achievement to determine readiness for the next unit achievement to determine readiness for the next grade or course achievement to determine academic readiness to begin college coursework achievement to determine academic readiness to begin career training coursework off/on/above track status for an outcome 2+ years out for planning/intervention
Programs & policies	Identify needs and track progress to develop/refine/evaluate… school policies/programs district policies/programs state policy/program implementation
Growth	Measure growth during a single marking period for policy/program evaluation

	Measure growth across 2+ marking periods for policy/program evaluation
	Isolate school effects on student growth for policy/program evaluation
	Isolate educator effects on student growth for educator evaluation

II. The State Layer

Introduction

The state-layer is well defined relative to the middle-layers of an assessment system, despite the added complexities of the three dimensional science standards. Specifically, virtually all states appear to be developing a single end-of-year assessment built around some variation of the NGSS performance expectations. The performance expectations are the basic unit of the standards and describe what students should be able to do. Each of the descriptions of what students should be able to do is based on the intersection of the three dimensions of the NGSS – the disciplinary core ideas, science and engineering practices, and the cross cutting concepts. To asses each performance expectation, again, virtually all states are developing item or task clusters aligned to the performance expectations (CCSSO & WestEd, 2015). These item clusters are being used in recognition that adequately assessing a performance expectation requires more than a single item. Following the *Science Assessment Item Collaborative Assessment Framework for the Next Generation Science Standards* (CCSSO & WestEd, 2015), item clusters are made up of multiple items based on at least one common stimulus.

Even with these commonalties among states, the assessments being used or developed still vary greatly in terms of how they capture performance in relation to three dimensional science standards. This section, like the following section on the district layer, focus on the design of the assessments within each level, the frameworks that defines the content of the assessments, the way in which student performance is modeled and the ways in assessment results are used. Unlike the following district-layer section, this work focuses less on the ways in which the decisions made at the state-layer can complement the district-layer, but instead simply attempts to capture what has been done currently. In doing so, this section attempts to capture the possible design decisions at the state-layer so that the district-layer can be defined *in relation* to the state-layer. In the best case, these two layers and all others are designed together, to better insure coherence. However, defining the state-layer and then designing the district-layer to complement it reflects the reality around assessment development and use – that the state-layer, and specifically the federally mandated state assessment, are often a given. Acknowledgement of this reality should not be taken as a sign that the district-layer should copy the state-layer, instead the district-layer should be carefully designed to address many of limitations of the state-layer.

Assessment Design

Here assessment design refers to the number of assessments given, the timing of those assessments and the distribution of content across assessments, both within and across grades. In terms of state assessments mandated by the Every Student Succeeds Act of 2015, this most often takes the form of end-of-year or end-of-course assessments that attempt to cover the domain, in this case the NGSS. One important question regarding the assessment framework is on the range of standards to assess.

Specifically, will the assessments be designed to assess grade-specific standards at each and every grade, or will the assessments be designed with grade band standards in mind? The later means that some standards will likely not be assessed, unless matrix sampling is used. Related to this is whether the assessments are given for each grade or course, or once within a grade-band. For example, a state might have a single assessment of 5th grade standards at the end of 5th grade to account for the elementary grade band. Alternatively, a state might incorporate a sampling of all of the elementary standards into that 5th grade assessment. A third option would be to have assessments of grade-level standards within each and every grade.

Also, it is worth noting that the end-of-year single summative assessment is not the only design that could be used for federally mandated state-level assessments. A state-level set of assessments could be developed in which the assessment content is partitioned across multiple smaller assessments. The item-cluster or task-cluster based design of NGSS assessments does easily lend itself to more distributed forms of assessment like those discussed within the middle-layers section. Such approaches could be viable under the interim option of the Every Student Succeeds Act[9] (ESSA, §1111(b)(2)(B)(viii) or the innovative assessment pilot option of ESSA (§1204) , assuming the state is willing to conduct a pilot with a limited number of districts. Finally, a state may wish to develop a more modularized or otherwise novel set of assessments and provide these to schools and districts for use as they see fit – separate from the assessments used to meet federal requirements. Such a set of assessments could be as simple as a task bank or more complex, in the form of more modular assessments. In reading and math, examples of these assessments include the modular and interim assessments provided by Wyoming[10] and the interim assessment blocks provided by the Smarter Balanced assessment consortia. In science, Kentucky provides classroom embedded assessments and through course assessments.

Assessment Design

- Generally end of year summative assessment, although states may also develop assessments for district use.
- Possibilities for standards coverage include:
  - ○ Grade Level and End of Course[1]
  - ○ Grade Bands
- Administered once per:
  - ○ Grade
  - ○ Grade Band

These boxes are meant to provide a quick summary of possible decisions, but are by no means comprehensive.

[1] Assuming the state has differentiated the NGSS by grade or course in middle- and high-school.

Assessment Framework

What, exactly the standards are varies from state to state. Some states, like Michigan and Wisconsin, have adopted just the performance expectations as the state standards and have therefore excluded the foundational boxes, which explicitly define the disciplinary core ideas, science and engineering practices, and the cross cutting concepts that define the performance expectations, in the state definition. Other states, like Washington and New Jersey, have adopted both the performance expectations and the foundational boxes, in more strict adoption of the NGSS. Taking a slightly different

[9] See Dadey & Gong (2017) for a more detailed discussion of the ESSA interim option.

[10] See https://edu.wyoming.gov/educators/state-assessment/wy-topp/ and https://edu.wyoming.gov/downloads/assessments/2015/2015ATFreport.pdf

approach than these prior states, states like Nebraska and Delaware view the performance expectations as valuable examples of intersections of the disciplinary core ideas, science and engineering practices, and the cross cutting concepts – but not the only intersections. Thus any combination of the DCIs, SEPs and CCCs is considered part of the standards.

Even after the standards are adopted, there are a number of choices that impact assessment development and the subsequent interpretations of assessment results. Key choices involve whether any aspect of the standards is foregrounded and how the content boundaries of the standards are defined. These two choices help make explicit the values expressed in the development of NGSS standards, and ideally, these choices should directly related to the overall claim to be made based on the assessment results, any related sub-claims and the achievement level descriptors. In an ideal assessment development process, the claims, sub-claims and achievement level descriptors (ALDs) would be clearly expressed before the development process began, thus making clear the extent to which aspects of the standards have been foregrounded and what the boundaries are. In this ideal process, the overall claim, sub-claims and ALDs would all be developed with a clearly articulated intended use in mind, so that the evidence elicited by the assessment would better support that use. However, given the impetus to start at the classroom level and the complexity of the NGSS, development of state-level assessments have often jumped directly to the development of item clusters, without first developing the overall claims, sub-claims and ALDs. Consequently, these choices may only be evident after the re-inspection of the assessment and its development process.

Assessment Framework

- Standards defined as the:
  - Performance Expectations
  - Performance Expectations + Foundational Boxes
  - Performance Expectations + Foundational Boxes + Other Intersections
- What aspects of the standards are foregrounded, if any:
  - A dimension
  - Transfer or problem solving
- What is assessed each year:
  - A fixed subset of the Performance Expectations
  - A subset of the Performance Expectations matrix sampled across years
  - All the Performance Expectations via matrix-sampling within year
- Are item clusters based on
  - A single Performance Expectation
  - A Performance Expectation bundle, that may or may not preserve the pairings of dimensions

Foregrounding. The NGSS seems somewhat amorphous, in that the same federally mandated state assessment could be reported in different ways, given the three dimensional standards. For example, the results of an assessment could be reported in terms of sub-scores based on the disciplinary core idea domains[11] – possibly diminishing the importance of the science and engineering practices or the crosscutting concepts. Groupings of the science and engineering practices or the crosscutting concepts could be used in similar ways. As an example, the science and engineering practices could be grouped into Investigation, Sensemaking and Critiquing (cf., McNeill, Katsh-Singer & Pelletier, 2015). A similar

[11] These domains are Physical Science, Life Science, and Earth and Space Science, as well as Engineering, Technology, and Applications of Science.

example for the crosscutting concepts is be causality, systems and patterns (cf., Moulding, 2012). Restated, the assessment blueprint could be organized in different ways to emphasize each dimension, possibly at the expense of the other dimensions. Finally, this foregrounding can be expressed through claims, subclaims, performance level descriptors, or any combination of the three. One common approach appears to be ALDs that are developed to each the disciplinary core idea domains, resulting in "multidimensional" ALDs. However, it is an open question as to whether and how each dimension of these ALDs interact with one another to define student performance. Should, for example, these ALDs be rooted in empirical analysis, if say, it is found that one of the disciplinary core idea domains is easier for students than others? Such questions currently do not have good answers.

Foregrounding, however, can also go beyond simply re-expressing the assessment in terms of one of the dimensions. The assessment could also be structured such that one dimension is more central. In this case, foregrounding is likely synonymous with developing an assessment claim that is based on a clearly articulated vision of science learning. For example, the science and engineering practices might be emphasized by creating item clusters that draw on a far wider range of practices than implied by any particular performance expectation. In this example, each item cluster might represent a substantial chunk of a scientific investigation. This appears to be the approach Kentucky has taken in the development of their NGSS based assessments. In addition, foregrounding is not limited solely to the NGSS dimensions. One alternative is to foreground the ability to solve new problems or understand novel phenomenon by drawing on the DCIs, SEPs and CCCs captured in one or more performance expectations or additional intersections of the dimensions.

Content Boundaries.

Even after defining what, exactly, the standards are, there are still a number of choices in translating the performance expectations to an actual assessment. Defining the content boundaries, or equivalently, defining much of the assessment and item specifications, is needed. Most concretely, there is a question as to whether the assessment will try to assess all of the PEs or a specific subset of the performance expectations. Assessing all of the performance expectations is only possible with a very lengthy assessment or via matrix sampling within or across years. Matrix sampling within year means that different students will take different item clusters aligned to different performance expectations, but across students at a given level (e.g. school) all the performance expectations are covered. Matrix sampling across years means that within any given year only a subset of the performance expectations will be covered, but across years all performance expectations are covered. Alternatively, a specific subset of performance expectations could be used to define the assessment, either as a fixed form or as the basis for within or across year matrix sampling. In addition, performance expectations may be treated "as-is" meaning there is a one to one mapping between an item cluster and a performance expectations, or "bundled," meaning that two or more performance expectations are combined into a single item cluster (CCSSO & WestEd, 2015). Then items within the bundle are either aligned to the dimensions with a single performance expectation, or the collective set of dimensions across the performance expectations within the bundle, in a type of "mix and match" approach.

Dimensionality

The ultimate aim of implementing a measurement model is to develop and maintain a reporting scale or scales that can support the claim(s) and sub-claims to be made based on the assessment results.

Dimensionality

- Unidimensional IRT
- Multidimensional IRT
  - Bifactor Model with secondary dimensions based on item clusters or standards specific information

Even when the claims are created early on in the assessment development process and a principled framework like evidence centered design is applied, the choice of measurement model is not always clear. Complicating considerations of modeling is that an item cluster approach adopted by virtually every state developing a federally mandated state assessment of the NGSS may not work well with the typically used unidimensional item response theory (IRT) methods. Simply put, the problem is that student performance on an assessment may not be summarized well[12] using a single scale (i.e., a single latent variable, generally referred to as ability or theta). Specifically, there are likely "cluster effects," meaning that after accounting for variability in student performance across item clusters, there is still a substantial amount of variability that is left over and associated with each and every cluster[13]. Such variability can be problematic, and may result in biased[14] item and student parameters, as well as negatively biased standard errors – assuming the bifactor model is true (see Wainer, Bradlow & Wang, 2007 for a general treatment and Rijmen, Turhan & Jiang, 2018 for an application within the context of the NGSS).

On some state NGSS assessments that incorporate item clusters, the exact amount of variability appears to differ by item cluster, but the general magnitude of the item cluster specific variance is large enough to suggest that student performance is not well summarized by a single scale[15] (e.g., Rijmen, Turhan & Jiang, 2018). If a single scale cannot describe student performance well, then two interrelated questions arise: (1) what scales would summarize student performance well (i.e., what latent variables should be used)?, and (2) what interpretations can be provided for these scales? A prevailing approach is to apply a specific type of multi-dimensional item response theory model, the bifactor model (or almost equivalently a testlet response model, see Wainer, Bradlow & Wang, 2007), to account for the unique cluster variance. Within this model, student performance on any item within an item cluster is summarized by an overall or general scale (i.e., a general latent variable) as well as a cluster specific scale (i.e., a cluster or secondary variable; one for each cluster). These cluster variables account for the unique cluster variance. In usual applications of the bifactor model, the general scale is primary and the cluster specific scales are ignored. Rijmen et al.'s (2018) application of the bifactor model is no different. Essentially, the unique item cluster variance is discarded and only the overall scale is reported. Whether this is consistent with the claim to be made about students is an open question. Item clusters are often

[12] That is, the assessment data are essentially unidimensional (see Stout, 1987).

[13] I.e., the item responses display local item independence when examined from a unidimensional model.

[14] Here the term bias is used in the statistical sense, meaning that the expected value of the parameter in question is systematically different than the true value.

[15] These results are based on assessments that score students not on items, but on "scoring assertions" which are derived not only from the answers a student provides to an item, but also their interaction with the stimuli of an item cluster (as captured through a digital assessment platform). There are often many more scoring assertions than items, meaning that the scoring assertion approach may be creating larger cluster effects than would be seen if just item response were used.

built with the idea that students are using the three dimensions to solve a problem or make sense of a phenomenon. Thus some of the unique item cluster variance that is being discarded is likely related to this problem solving or sense making. On the other hand, one might argue that a student's ability to solve problems or make sense across multiple problems or phenomena is of interest.

Defining the secondary variables based on clusters is not the only way to define a bifactor model, nor is it the only model[16] that could be used to summarize student performance. The secondary variables of the bifactor model can be based on any grouping of items. However, this raises a question about whether that grouping can account for the additional variance not accounted for by the general factor. Such groupings could be uncovered through exploratory analysis or defined based on some aspect of the standards – for example the DCI domains, as suggested by Martineau (2018, February).

Use Cases

For any assessment, the developers must clearly articulate ways in which the assessment results are meant to be used and ways in which the assessment results are not meant to be used. Specificity in these uses are key. All too often state-level assessments are designed to provide a school-level percent proficient as one part of a state's system of school identification for targeted or comprehensive support, but are also ascribed vague classroom relevant uses like, "to improve teaching and learning". If such uses are put forth, they should be made specific. Table 1 provides a starting point for such specificity, but it is far from complete. One way to structure such specificity is through a theory of action that couches the use of assessment results in terms of outcomes and the mechanisms which produce the intended effects of an assessment system. The theory of action should also specify the action mechanisms by which the effects are supposed to be caused, which in turn dictates the interpretations that an assessment or assessments will need to support for examples of theories of action applied to assessment systems, see Bennett, 2010; Bennett, Kane & Bridgeman, 2011; Hall, Domaleski, Russell & Pinsonneaul, 2016). Ideally, the theory of action would encompass multiple layers of the system of assessments, such that the uses of assessment work together to achieve the intended outcomes.

In terms of the broad categories of use outlined in Table 1, the use of federally mandated state-level assessments often fall into the category of "evaluate schools,", although many states have opted to not include science as part of their ESSA compliant accountability system. Many other uses of a state-level assessment could be applicable – within the context of a supporting theory of action. For example, a district might use state-level assessment results to examine what grade-levels have the lowest percentage of students at proficiency and then target professional development to teachers in those grades or provide supplemental support to the cohort of students corresponding to that grade. A

[16] Another, albeit unexplored, approach might be to instead treat the item clusters in the same way that Essential Elements (EE) are treated in the Dynamic Learning Maps (DLM) assessment system. That is, for each item cluster, a latent class model could be applied to estimate the probability of mastering the performance expectation. However, the DLM system does have five levels within each EE. Adapting this to the NGSS would mean that each PE or PE bundle would need to have multiple levels of performance articulated, perhaps through use of the NGSS Evidence Statements.

guiding question in examining use is whether the results from the state-level assessment(s) alone are sufficient to support the intended use. The results of the last twenty years of accountability policy would suggest that state-level information alone is not sufficient and that information from other layers of the system should be considered.

III. The District Layer

Introduction

The layers of an assessment system between the classroom and the state, and the district layer in particular, is where a great deal of policy intention and practical reality collide. This section focuses solely on the district level, but the commentary could apply at any other level between the classroom and state, be it school, region or some other unit. This focus stems from the idea that the district layer fits the "goldilocks" principle – it is close enough to classrooms to tie to curriculum and instruction, but far enough away to leverage economies of scale to develop or purchase assessments. Importantly, districts play a key role in curriculum development, which should be completed by assessment. Moreover, local control is king in United States. Thus the district layer and other middle layers are a key locus of control in the development of an assessment system (cf., Marion, 2018). Therefore, states should work to flexibly support district-level assessment development, but also consider ways in which the state-level assessment(s) can complement the district-level efforts. Wyoming, for example, has developed state-level optional interim and modular assessments in mathematics and reading to be used at the classroom, school and district levels. The state does not presuppose a given use, but rather sees the assessments as tools that can be use flexibly to support a possible range of uses, at the users discretion.

Also, perhaps more than any other layer, the purposes of assessments are conflated within the district-layer. For example, district-level assessments used for monitoring purposes are also touted as having relevance to teaching and learning, when, at best, these assessments are distal measures of learning with loose and undefined connections to classroom instruction. The assessment audit provision of the Every Student Succeeds Act can be seen as a reaction to this disconnect, as well as to the proliferation of district assessments. This proliferation has often meant that students are taking many different district assessments that are duplicative or extraneous in purpose. The distinction between classroom and monitoring assessments (NRC, 2014) is particularly useful here. Specifically, classroom uses of assessment need to be "fire-walled" from monitoring uses, as monitoring uses almost always swamp classroom uses (cf., Campbell, 1979).

Before turning to issues on assessment design, assessment frameworks, modeling considerations and use cases, it is worth pointing out that there is no clear distinction between the issues discussed above in terms of the state-layer and those discussed next in relation to the district-layer. Depending on the use of an assessment, any of the issues could apply to any layer. However, state assessments are often used for monitoring in compliance with federal accountability policy and district assessments are often used for monitoring of programs and educators (see Shepard et al., 2018, p. 26) and, possibly, for classroom purposes. Thus the issues discussed do fit somewhat naturally at the given layers. Regardless of the sharp distinction between the state- and district-layers, or lack thereof, the idea of complementarity is across layers is key.

Assessment Design

One model for the development of district assessments is for the state to create assessments that are then adopted by interested districts. Alternatively, assessments might be developed by districts[17], potentially with the support of the stat,e or purchased by districts from vendors. Any one of these approaches has the potential to produce assessments that do not work well together within the district-layer, or across the district- and state-layer. The processes involved in joint district and state assessment development may help safeguard against such lack of complementarity, but this is not a given. For many uses within the district-layer, curricular specificity is paramount. Being able to adapt assessments to local curriculum is then very important, and sometimes difficult to do with state supplied assessments and many off the shelf assessments[18].

Assessment Design

- Level of Modularity
  - ○ Fixed Designs that broadly measure the domain generally via a single assessment
  - ○ Block Designs, involving multiple assessments each aligned to a large chunk of the domain
  - ○ Modular Designs, involving multiple assessments each aligned to a fine grained part of the domain
- Are the assessment(s)
  - ○ Administered on demand
  - ○ Held securely
  - ○ Part of a digital assessment platform

Some assessment designs will likely support adaptation to local curriculum more than others. Specifically, *block* and *modular* designs are more easily adapted to match multiple scopes and sequences of instruction than are *fixed* designs. These types of designs have been discussed previously (see Gong, 2010), but the distinction of fixed, block and modular designs is relatively novel to this work[19].

Fixed designs include those assessments that broadly measure a given content domain, generally through a single assessment that surveys the domain. This design is commonly used in interim assessments. Such assessments are often deemed to have a “summative” design, meaning that their blueprints reflect those of the state-level summative assessment (see Gong, 2010 for more on these types of designs). A variation on this summative fixed design is one in which the assessment follows the same blueprint as the state-level summative assessment, but is shorter in length. When used for summative purposes, these shorter assessments are sometimes referred to as “mini-summatives.” Block

[17] One interesting approach to district development might be to create consortia of districts interested in pooling resources to develop assessments.

[18] Although some commercial assessments are packaged alongside curriculum and *may* bypass some or many issues related to curriculum sensitivity.

[19] The term “modular” has been used previously, for example in the Competitive Preference Priority 1 of the now closed Enhanced Assessment Grants Program (U.S. Department of Education, 2016). The grant program called for “approaches to transform traditional, end-of-year summative assessment forms with many items into a series of modular assessment forms, each with fewer items than the end-of year summative assessment” (p. 5)

designs involve multiple assessments, each of which measures large chunks of the content domain. To completely measure the domain multiple assessments need to be administered, often at the discretion of the end users. The order and timing of the administration of block assessments therefore varies based on the needs of those administering them. The Smarter Balanced interim assessment blocks are an example of this design. Delaware's optional unit based assessments are another. Modular designs also involve multiple assessments, but under this design these assessments measure fine grained parts of the content domain. One example of a modular design is a bank of item clusters aligned to individual performance expectations, like the one provided by Kentucky in a through course task bank. The individual assessments within such banks, be they items, item clusters or other combinations of items, could be use individually or combined to make larger forms that reflect local scopes and sequences of instruction. Recommended combinations could be compiled by a district or state to match various curricula throughout the state. Such supports are likely needed, as full modular designs likely require additional support to help users know how to meaningfully group, interpret and use the assessments.

Implementing any of above designs involves addressing a number of concerns that are outside the scope of this work. Such concerns include whether the assessments with the district-layer (a) can be administered on demand or are provided within fixed windows, (b) are open and can thus be seen by the public or are held securely, (c) are part of a digital assessment platform that provides automatic scoring or the ability to input student responses on hand scored items or tasks.

Assessment Framework

At the district-layer the choice of what to assess is far less constrained that at the state-layer. A pressing question is how the district-layer can best work with the state-layer for a given use. Often, the default answer to this question is not at all – that is, the district layer has been purposely made to work in isolation from the state layer. In other cases, the district-layer assessments are meant to show progress towards, or predict performance on, the federally mandated, end of year state assessment (e.g., Perie et al., 2009). These two cases a very narrow subset of the ways in which the relationships between the state- and district-layers can be defined. Part and parcel of these relationships is how the content of the district-layer assessments are defined, and defined in relation to the state-layer assessments.

Assessment Framework

- Is the approach to defining what is assessed intentionally different or the same as the state-level?

An important distinction is whether the district-layer assessments will take the *same* approach to measuring the domain as the state-layer, in terms of both depth and breadth, or a *different* approach? In particular, will the assessment framework of the district-layer assessments mirror that of the state-level? There are viable reasons for both approaches. For example, a district might choose to assess intersections outside of the performance expectations assessed on the state-level assessment(s) to signal the breath of the NGSS, more fully represent the standards and thus attempt to mitigate any narrowing of the curriculum. Or district assessments might be used to measure the performance expectations "as is" while the state-level assessment focuses on transfer, as demonstrated by the student's ability to solve novel problems or understand novel phenomenon. This later approach is the basis of the Delaware design. An additional example is one in which the district assessments measure the domain in a cumulative fashion, in which the content of a set of assessments becomes increasingly large until the domain is covered. Variations in assessment frameworks are explored in detail in Gong (2010), as well as in Herman (2017) and Brookhart (2013). In addition, it is worth noting that certain the assessment framework approaches dovetail with certain assessment design approaches. For example, the cumulative assessment framework approach mentioned as an example above would likely work with the block design and less so with the fixed design.

Dimensionality

A guiding question is whether classical test theory is sufficient (i.e., total scores are all that is needed) or if more complex models are need, like IRT models. Many uses of district-layer assessment results do not require such complex measurement models. For example, more complex measurement models are likely not needed if an assessment is being used across schools by a group of teachers instructing the same content, for the purpose of determining whether the overall scope and sequence of instruction should be shifted. Other uses, however, may require more complex models. For example, tracking trends across years using different assessment forms generally requires the use of a model like IRT. That is, models like IRT support the comparability of results, so that comparisons can be made more readily across time or units. Uses that require comparisons across schools, districts or years likely necessitate a measurement model – particularity if the item clusters or assessments change across students.

Dimensionality

- Will the model be the same or different as the state-level model?
- How will data from an adequate sample of students be collected?

However, applying a measurement model to district assessments is not necessarily more complicated than applying these models at the state-level, even if a more modular design is desired (see Dadey, Tao & Keng, 2018 for an example in mathematics). The same concerns that arise at the state level are likely to also arise at the district level, including those about multidimensionality. Rather, a more difficult problem with the application of the measurement model is how to create an adequate sample of students from which to gather data. An equally difficult, but more conceptual problem is about what kind of score is to be reported, and how it relates to the state-level. Specifically, if a scale score or classification is desired, should it be reported using the same scale, and thus underlying measurement model, as from state-level assessments, or should it differ (i.e., is there a desire for comparability in scores across levels)? The Smarter Balanced interim assessment blocks report classifications that are the same as those provided on the Smarter Balanced summative assessment (i.e., the same achievement levels based on the Smarter Balanced summative measurement model). This model appears to also be adopted by some vendors, who provide open access item clusters based on a measurement model created using summative data. On the other hand, there are also reasons for not doing so (see the example of district-layer assessments of learning progressions presented by Briggs, 2018)

Use Cases

As noted previously, the uses of assessment results are often conflated at the district level – with vague uses related to instruction tied to very specific monitoring purposes. This is not to say that the district-level assessments cannot be relevant to teaching and learning, particularly when a design can be tailored to instructional sequences and there are supports to do so. However, care does need to be taken to insure that such uses are actually supported by the assessment results.

With the right design and accompanying theory of action, any and all of the uses detailed in Table 1 could be supported by district-level assessment. Any use that is evaluative in nature (i.e., used for monitoring) likely needs to be separated from classroom use. For example, if a fully modular assessment item cluster bank is meant to be used for monitoring purposes, some tasks could be held securely apart from those used for classroom purposes. An alternative approach, might be to simply limit the types of evaluations to low stakes and provide open access to the assessments.

Finally, the uses across the state and district-levels should be examined for possible connections, or lack thereof. While some uses naturally draw on both levels (e.g., early warning uses), other uses may be completely disconnected. In addition, some designs are better for particular uses than others (e.g., program evaluation would not likely be based on item clusters from a bank, although such hurdles could be overcome).

IV. Tying the State and District Together: Complementary Design

Pellegrino et al. (2001) outline three criteria for evaluating the quality of an assessment system. These criteria are comprehensiveness, coherence and continuity. Comprehensiveness means that multiple sources of information collected through difference processes (e.g., standardized assessments, performance assessments) are considered. Coherence means that the models of student learning underlying all of the assessments are compatible and that alignment is needed among curriculum, instruction, and assessment. Finally, continuity means that assessments should measure student progress in relation to a model of student progression.

I suggest that any combination of these criteria can be used as a guide to improve the overall quality of an assessment system, and should be considered in light of the categories that I have used to characterize each layer of the assessment system (assessment design, assessment framework, modeling considerations and use cases). Not all of the three criteria map onto each of the categories well, but considering the interplay is useful. For example, introducing item clusters that are performance based to improve comprehensiveness could better support a variety of uses, and would likely make up for some of the limitations at the state-level. Developing modular district-level assessments, with recommended administration patterns based on common curriculum, could improve coherence.

References

Abrams L. M., McMillan J. H. & Wetzel, A. P. (2015). *Implementing benchmark testing for formative purposes: teacher voices about what works*. *Educational Assessment, Evaluation and Accountability, 27*(4), 347-375.

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., Kane, M. and Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Paper presented at the 2011 Invitational Research Symposium on Through-Course Summative Assessments: Atlanta, GA.

Briggs, D.C. (2018, Feburary). *From learning progressions to multidimensional models (and back)*. Paper presented at the Winter Meeting of the Science State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.

Brookhart, S. M. (2013). Comprehensive assessment systems in service of learning: getting the balance right. In R. L. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 165-184). Charlotte, NC: Information Age Publishing, Inc.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: the role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education, 85*(2), 186-204.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Educational and Program Planning, 2*(1), 67-90.

CCSSO & WestEd. (2015). *Science assessment item collaborative assessment framework for the next generation science standards*. Washington, DC: The Council of Chief State School Officers.

Conley, D. T. (2018). *The promise and practice of next generation assessment*. Cambridge, MA: Harvard University Press.

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education, 112*(4), 469-495.

Clune, W.H., & White, P.A. (2008). *Policy effectiveness of interim assessments in providence public schools*. WCER Working Paper No. 2008-10, Wisconsin Center for Education Research. Madison: University of Wisconsin.

Dadey, N., & Gong, B. (2017, April). *Using interim assessments in place of summative assessments*? Consideration of an ESSA option. Washington, DC: Council of Chief State School Officers. Available online: https://www.nciea.org/sites/default/files/inline-files/ASR%20ESSA%20Interim%20 Considerations-April%202017.pdf

Dadey, N. Tao, S. & Keng, L. (2018, April). *Developing scale scores and cut scores for on-demand assessments of individual standards*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Diaz-Dilello, E. K. (2011). *A validity study of interim assessments in an urban school district* (Doctoral Dissertation). Available online at: https://scholar.colorado.edu/educ_gradetds/10/

Darling-Hammond, L., Pecheone, R., Jaquith, A., Schultz, S., Walker, L., & Wei, R.C. (2010). *Developing an Internationally Comparable Balanced Assessment System That Supports High-Quality Learning*. Presented at the National Conference on Next Generation K-12 Assessment Systems, March 2010, Washington, D.C. Available online at: http://www.k12center.org/rsc/pdf/Darling-HammondPechoneSystemModel.pdf

Davidson, K.L., & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments*. CRESST Report 806. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Gong, B. (2010). *Some implications of the design of balanced assessment systems for the evaluation of the technical quality of assessments*. Dover, NH: The National Center for the Improvement of Educational Assessment, Inc. Available online at: https://www.nciea.org/sites/default/files/publications/RILS_Gong2010.pdf

Gong, B. & Norris, M. (2018). *Thinking about claims for assessments of the Next Generation Science Standards: Domain definition, performance level descriptors, and reporting categories.* Dover, NH: The National Center for the Improvement of Educational Assessment, Inc.

Gibbons R. D., & Hedeker D. R. (1992) Full-information item bi-factor analysis. *Psychometrika*, 57, 423–436.

Hall, E., Domaleski, C., Russell, M. & Pinsonneault, L. (2016). A framework to support accountability evaluation. Dover, NH: The National Center for the Improvement of Educational Assessment, Inc. Available online at: https://www.nciea.org/sites/default/files/pubs-tmp/A_Framework_to_Support_Accountability_Evaluation.pdf

Herman, J. (2010). *Coherence: Key to next generation assessment success* (AACC Report). Los Angeles, CA: University of California. Available online from www.cse.ucla.edu/products/policy/coherence_v6.pdf

Herman, J. (2017). *Interim Assessments in brief.* Los Angeles, CA: University of California. Available online from http://www.csai-online.org/sites/default/files/resources/4666/InterimAssessmentsBrief.pdf

Marion, S. (2018). The Opportunities and Challenges of a Systems Approach to Assessment. *Educational Measurement: Issues & Practice, 37*(I), pp. 45-48.

Martineau, J. A. (2018, February). *The intersection of measurement model, equating, and the Next Generation Science Standards*. Paper presented at the Winter Meeting of the Science State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.

Martineau, J., Dewsbury-White, K., Roeber, E., Vorenkamp, E., Snead, S., Flukes, J. (2018). District Assessment System Design Toolkit. Dover, NH: National Center for the Improvement of Educational Assessment (Center for Assessment). Available at https://www.nciea.org/dasd-toolkit/.

McNeill, K. L., Katsh-Singer, R. & Pelletier, P. (2015). Assessing science practices – Moving your class along a continuum. *Science Scope, 39*(4), 21-28.

Moulding, B. (2012). *Science and engineering practices organized around gathering, reasoning, and communicating*. Available online at: http://www.csss-science.org/downloads/bcsse/denver/BrettMouldingGatheringReasoningCommunicating.pdf

Penuel, W. R., Frumin, K., Van Horne, K., & Jacobs, J. K. (2018, April). *A phenomenon-based assessment system for three-dimensional science standards: Why do we need it and what can it look like in practice?* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY. Available online at: http://learndbir.org/resources/A-Phenomenon-based-Assessment-System-for-Three-dimensional-Science-Standards.pdf

Ryan, J. M. (2010). Envisioning a state educational system: Improving learning through a comprehensive assessment system. Olympia, WA: Office of Superintendent of Public Instruction.

Perie, M., Marion, M., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28* (3), 5-13.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment*, *92*(6), 544–559. http://doi.org/10.1080/00223891.2010.496477

Rijmen, F., Turhan, A., & Jiang, T. (2018, April). *An item response theory model for next generation of science standards assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Stout. W. F. (1987) A non parametric approach for assessing latent trait unidimensionality. *Psychometrika,* 52, 589-617

Shepard, L., Davidson, K., & Bowman, R. (2011). *How middle-school mathematics teachers use interim and benchmark assessment data* (CRESST Report 807). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shepard, L. A., Penuel, W. R., & Pellegrino, J. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues & Practice, 37*(1), pp. 21-34.

Supovitz, J. A., & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematical use student performance data to guide improvement*. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

U.S. Department of Education. (2016). *Application for new grants under the Enhanced Assessment Instruments Grant Program (EAG)* (CFDA 84.368A). Available online: http://www2.ed.gov/programs/eag/eag2016application.pdf.

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