Using Principled Assessment Design to Support Formative Assessment and Students' Opportunities to Learn

Dr. Christina Schneider

NWEA

Paper presented in the NCME coordinated session: Opportunity to Learn: Impact on Large-Scale and Classroom Assessment Design and Interpretation on April 30, 2017, in San Antonio, Texas.

Introduction

Teachers need more sophisticated and nuanced support systems to understand and facilitate student learning. These supports go beyond state standards, district curriculums, and published textbook materials. Teachers need more information regarding how a state interprets its standards and what mastery of those standards represents on large scale assessments. They need supports in regard to how to monitor student learning over time in a way that focuses them not on the number of correct responses a child is providing, but rather, whether those correct responses represent more sophisticated reasoning and content acquisition than was observed previously. Teachers need support in using sophistication of reasoning evidence to take instructional actions.

When a teacher analyzes evidence of student learning and uses that information to either adjust instruction or provide feedback, he or she is using formative classroom assessment practice (Brookhart et al., 2008). The formative assessment theory of action is based in the idea that a teacher identifies where a child is, identifies what the child needs next, and the gap between where a child is and the learning target is closed. The action a teacher takes; however, can increase, decrease, or not influence student learning, depending upon if the action a teacher takes is the *right* action for a particular student. For these reasons, researchers have posited that formative assessment actions are not validated and accurate unless they result in increases in student learning (Brookhart, 2009; Nichols et al., 2009; Shepard, 2009).

Supporting the Foundational Assumptions of Formative Assessment

A teacher's ability to engage successfully in formative assessment practice is predicated on four important assumptions defined by Schneider and Gowan (2013). In this section the first three formative assessment assumptions are described (the fourth assumption is that students use information to close the learning gap), and evidence is provided to support these propositions.

Assumption One: Teachers Collect Accurate Information about Student Learning. Often teachers do not have sufficient information to accurately interpret the intended outcomes represented by state standards. As a result, they often use assessments that *in*accurately measure student learning of those standards. This proposition is supported by a research base spanning nearly 34 years. Researchers have historically found that many assessments teachers administer do not match rigorous state standards (Flemming & Chambers, 1983; Carter, 1984; Marso & Pigge, 1993). More recent evidence from Llosa (2005) and Yap et al, (2007) can help policy makers understand why. Llosa found teachers inconsistently interpreted standards with multiple parts. They ignored the parts of standards they did not understand, developed their own interpretation for the parts, or ignored the standard or parts entirely. Yap et al. found 34% of teachers in their study did not accurately interpret a state standard they self-selected as evidence to show their skill in this regard. Moss et al. (2013) found 50% of work assigned to students had minimal or no connection to a learning goal if the teacher was identified as struggling by their administrator. In such cases, the students are not being provided the opportunity to learn the intended content to the intended degree of rigor.

One reason that teachers may struggle to interpret state standards is because the standards represent, in words, children's actions and thinking within a particular content area. For example, states' use words to describe reading skills rather than provide teachers examples of students reading grade-level texts with accuracy, fluency, and comprehension, and states do not provide examples of student inferences and evidence a teacher can use to recognize "on track" student thinking. This may make the learning targets to which the state aspires ambiguous for a teacher. How well does a child have to analyze texts to be ready for success in the next grade?

An additional complication is that teachers may not recognize that within a standard can be ranges of content- and thinking-skill difficulty that describe different levels of achievement (Egan et al., 2012; Schneider & Egan, 2014). Moreover, not all standards are equal. Some standards are precursors to others; whereas, others represent more advanced levels of thinking with the content (Schneider & Johnson, in press). Individual standards measured in isolation, without a content connection to other within-grade standards, do not allow an individual teacher to determine what proficiency in grade-level content or "on track" performance represents. Being "on track" for success in the next grade is a content-centered process in which teachers define how much content and at what level of difficulty a student should be able to manipulate and problem solve with (Lewis et al., 2012; Ferrara & Lewis, 2012). This is a different notion from the common teacher belief that standards are objectives that teachers can check off. Being "on track" for success means not all content has to be answered correctly by a child, but that particular content at a particular level of difficulty, integrated with related standards, *must* be.

Assumption Two: Teachers Analyze Student Learning Evidence and Make Accurate Inferences. In addition to collecting accurate information about student learning, a teacher needs to accurately analyze and interpret the information collected. The scant research evidence about how often and how well teachers engage in this process suggests teachers rarely analyze student work at an individual student level. Ruiz-Primo et al. (2010) found that most teachers in their study did not analyze collected student work that was embedded into curriculum units for this purpose. When teachers do analyze student learning it tends to be at a holistic-classroom level using average test scores on assessments as the primary data tool (Schneider & Meyer, 2012; Hoover & Abrams, 2013).

The use of class-level means as the key data point for instructional actions may not be best formative practice. Information about the average child does not help a teacher diagnose gaps, confusions, or beauty in thinking for a single student (Schneider & Andrade, 2013). Moreover, for students who are ready to move on, such practice can restrict the range of content a child would have seen and has been shown to cause decreases in achievement over time (Schneider & Meyer, 2012). For students who need more time and practice to learn critical content, teacher decisions to move on can, over time, leave significant portions of the student population falling behind their peers.

Assumption Three: Teachers Provide Accurate Feedback and Instructional Adaptations. Determining the next instructional steps based upon assessment information is a highly complex teacher task (Heritage et al., 2009; Schneider & Gowan, 2013; Schneider et al., 2014), and evidence suggests roughly 30 to 40 percent of teachers need support in this area. Learning progressions can be a support for instructional actions because they offer likely instructional pathways. Such learning pathways can occur within a single standard (Schneider & Egan, 2014) and across standards within the course of a year.

Smith et al. (2006) defined a learning progression as the description of the increasingly more sophisticated ways of reasoning in the content domain that follow one another as a student learns. Clements and Sarama (2004) noted learning progressions (frequently referred to as learning trajectory in their work) describe levels of student thinking. Learning progressions must in the view of Furtak et al. (2014) not only describe how students learn, they must be an interpretive aid in analyzing that information, and a support for using the information for action. Assessment developers who comingle achievement level descriptors with the notions of learning progressions have the potential to provide a tool for teachers that support desired formative assessment practices: a systemically valid assessment that supports change in instruction and curriculum to foster the development of the cognitive skills that the assessment is designed to measure (Frederiksen & Collins, 1989).

Perie & Huff (2016) posited as educators work to personalize instruction centered on where a child is currently in their learning, test developers must begin creating assessments that provide guidance under such a framework. Test design centered in achievement level descriptors (ALDs) that show how students increase in their reasoning with specific content across achievement levels can support teachers in collecting accurate information about student learning, analyzing student learning evidence to make accurate inferences about what students know and can do (especially when coupled which some of the procedures underlying the ID matching method), and providing more targeted instructional adaptions based on the student's present level of performance. Under such an approach when test designers use principled approaches to test design, ALDs may be viewed as the interpretation regarding what a test score represents.

Principled Assessment Design based on Achievement Level Descriptors

Kane (2006) challenged assessment developers to explicitly

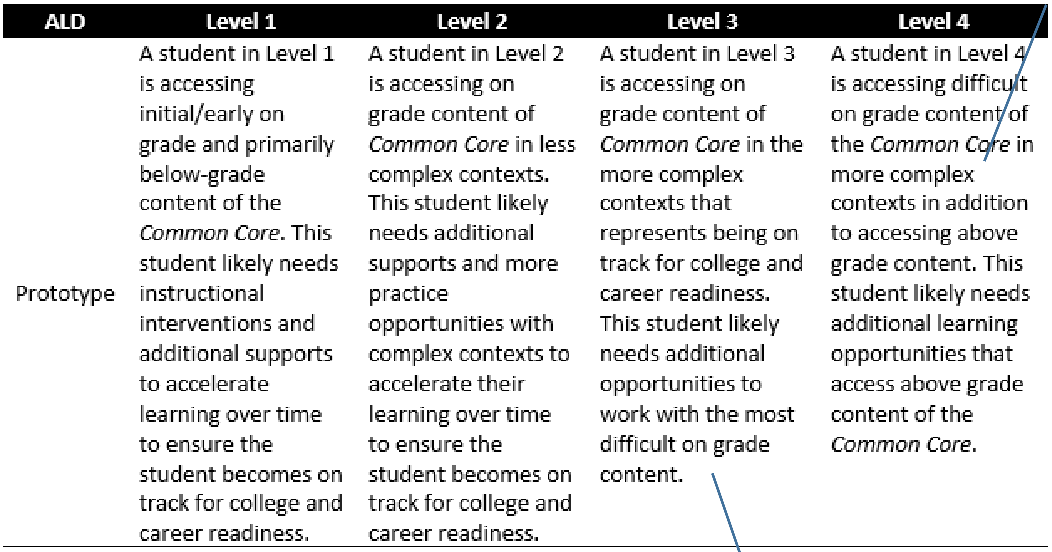
(1) define the intended interpretation of assessment scores,
(2) define the intended use of assessment scores,
(3) map the network of system inferences that lead from a student's performance on a series of tasks /items to the actions that are intended to occur based on the information, and
(4) map the network of system assumptions that lead from a student's performance on a series of tasks/items to the actions that are intended to occur based on the information.

Fundamentally, achievement level descriptors should be the linchpin of the interpretation and use argument for an assessment. Teachers and test developers should use the same types of evidence and standards interpretation to understand student learning. When states develop ALDs to articulate the observable evidence teachers and item developers should elicit to draw conclusions about a student's current level of performance, what that evidence looks like when students are in different stages of development represented by different achievement levels, and how the student is expected to grow in reasoning and content skill acquisition across achievement levels within and across grades, they better communicate how standards are interpreted for assessment purposes, how tasks can align to a standard but not be of sufficient difficulty and depth to represent mastery, and what growth on the test score continuum represents.

Policy ALDs - Claims

In the first stage of the ALD development framework (Egan et al., 2012), a state should develop Policy ALDs. Policy ALDs are currently important communication devices for the vision of intended test score interpretations and they hold the potential to explicitly define the primary intended use of test scores at a system level. For example, often a test score is intended to represent where a student is in their learning regarding the Common Core State Standards (CCSS). That is, students are in a stage of learning within the CCSS that is defined by the achievement level for grade-level content. The sample policy descriptor shown in Table 1 is intended to set the tone that students in more advanced achievement levels demonstrate content understanding in more complex contexts such as higher levels of Webb's DOK framework (Webb, 2005). When carefully crafted, they can be viewed as the assessment claim because they set the tone for how the content and cognitive demand is intended to be articulated along the test scale.

Table 1: Prototype Policy Descriptions



ALD	Level 1	Level 2	Level 3	Level 4
Prototype	A student in Level 1 is accessing initial/early on grade and primarily below-grade content of the *Common Core*. This student likely needs instructional interventions and additional supports to accelerate learning over time to ensure the student becomes on track for college and career readiness.	A student in Level 2 is accessing on grade content of *Common Core* in less complex contexts. This student likely needs additional supports and more practice opportunities with complex contexts to accelerate their learning over time to ensure the student becomes on track for college and career readiness.	A student in Level 3 is accessing on grade content of *Common Core* in the more complex contexts that represents being on track for college and career readiness. This student likely needs additional opportunities to work with the most difficult on grade content.	A student in Level 4 is accessing difficult on grade content of the *Common Core* in more complex contexts in addition to accessing above grade content. This student likely needs additional learning opportunities that access above grade content of the *Common Core*.

Claim: What the test construct is intended to represent.

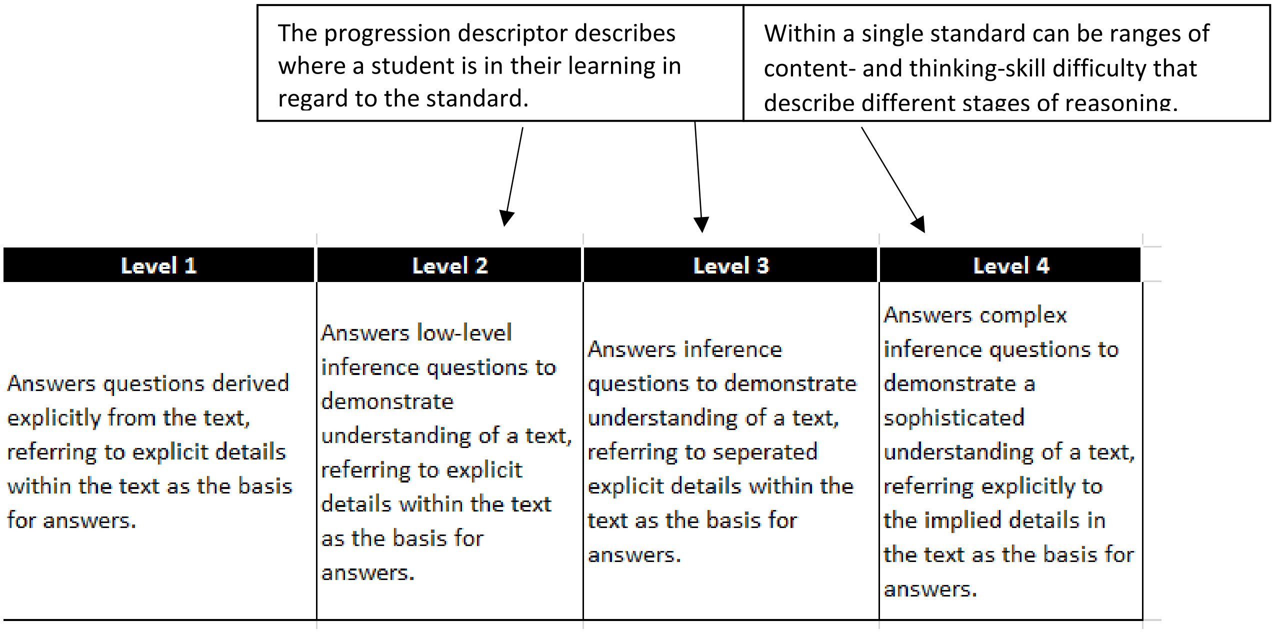
Intended use is instructional action.

Range ALDs – Evidence to be Collected

For each standard and achievement level on an assessment, Range ALDs should explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated across achievement levels. Schneider et al. (2012) wrote that for ALDs to be the foundation of test score interpretation, they should reflect more complex knowledge, skills, and abilities (KSAs) as the achievement levels increase (e.g., more complex KSAs should be expected for Level 4 than for Level 3). This notion is consistent with what is termed a learning progression or learning trajectory in the research literature.

Learning progressions are increasingly described in the literature as theoretical underpinnings for curriculum development, instruction, and assessment of learning. The purpose of a learning progression is to inform researchers and educators about general developmental pathways of learning so that they can set reasonable, achievable learning goals and provide appropriate guidance for instruction and assessment in each content area. Assessments that are designed to measure student growth as well as inform instruction should be derived from the combination of the learning goals and within-grade developmental progressions of those goals. Clements and Sarama (2004) wrote, "Developmental progressions . . . [are] descriptions of children's thinking and learning . . . and a related, conjectured route through a set of … tasks" (p. 83). The outcome of instructional tasks delivered in the classroom or assessment tasks delivered on a large-scale assessment should be the same: observable evidence of what students know and can do in relation to the stages of learning of on grade content from a state's standards.

Under a principled assessment design approach Range Achievement Level descriptors provided the intended content-based interpretations of what scale scores within a particular achievement level represent. They can provide teachers much needed information on ranges of content- and thinking-skill difficulty that can be found within a single standard as shown in Figure 1, and at times, across standards. The number of concepts and processes that students must integrate to respond to tasks must be monitored and made explicit in achievement level descriptors because this affects the difficulty of the tasks as well as state developed notions of mastery (Ferrara & Steedle, 2015; Schneider & Johnson, in press). Articulating how skills change and becomes more sophisticated across achievement levels, the observable evidence needed to support those conclusions, and what evidence of student work looks like from students who have "mastered" the standards is accomplished as one of the first steps of domain analysis found in principled approaches to test design such as ECD.



The progression descriptor describes where a student is in their learning in regard to the standard.

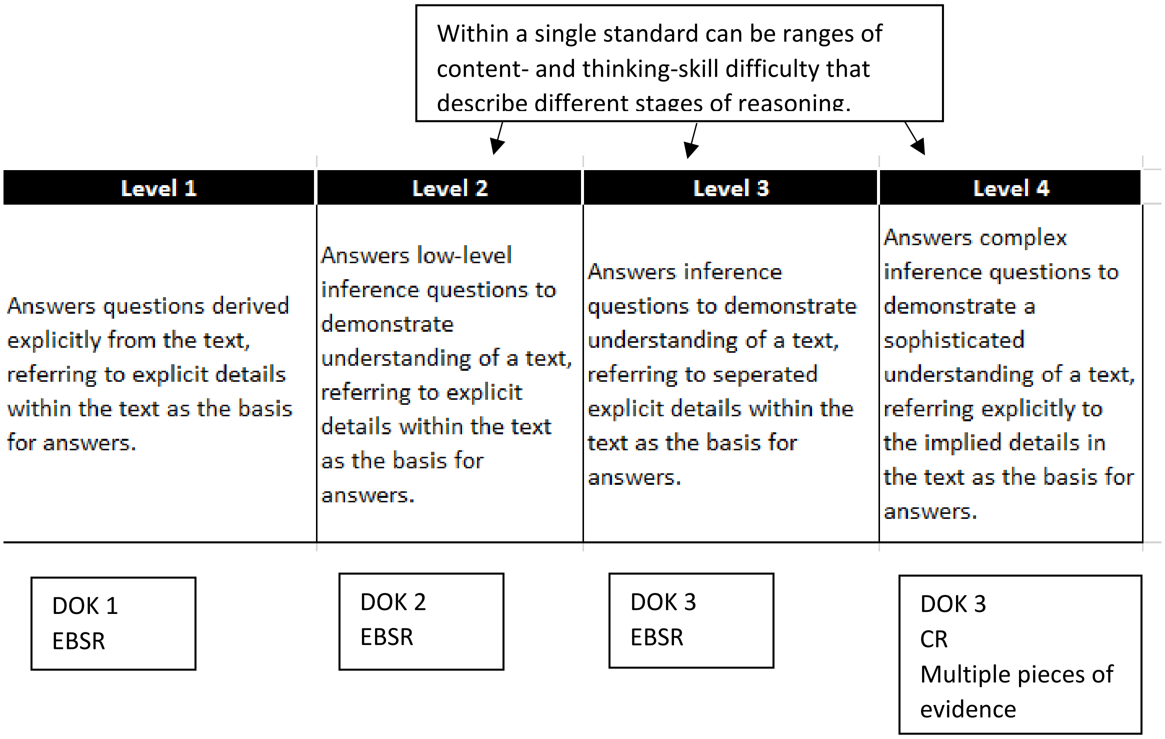
Within a single standard can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Level 1	Level 2	Level 3	Level 4
Answers questions derived explicitly from the text, referring to explicit details within the text as the basis for answers.	Answers low-level inference questions to demonstrate understanding of a text, referring to explicit details within the text as the basis for answers.	Answers inference questions to demonstrate understanding of a text, referring to seperated explicit details within the text as the basis for answers.	Answers complex inference questions to demonstrate a sophisticated understanding of a text, referring explicitly to the implied details in the text as the basis for answers.

Figure 1: Content Interpretation Defined in Range ALDS

Test Specifications and Item Specifications

As Range ALDs are developed based on how student thinking grows and the evidence needed to support that conclusion, Range ALDs have the opportunity to drive item development and test construction as test developers determine the item types that are optimally developed to support the intended score interpretations. As ALDs describe increases in student thinking and reasoning test developers have a rationale regarding why a percentage of particular item types (e.g., technology enhanced items, and constructed and extended response items) are necessary and the percentage of items that should be developed to particular levels of cognitive complexity within an item bank. Those decisions are driven based upon the construct-based evidence that should be collected as shown in Figure 2.



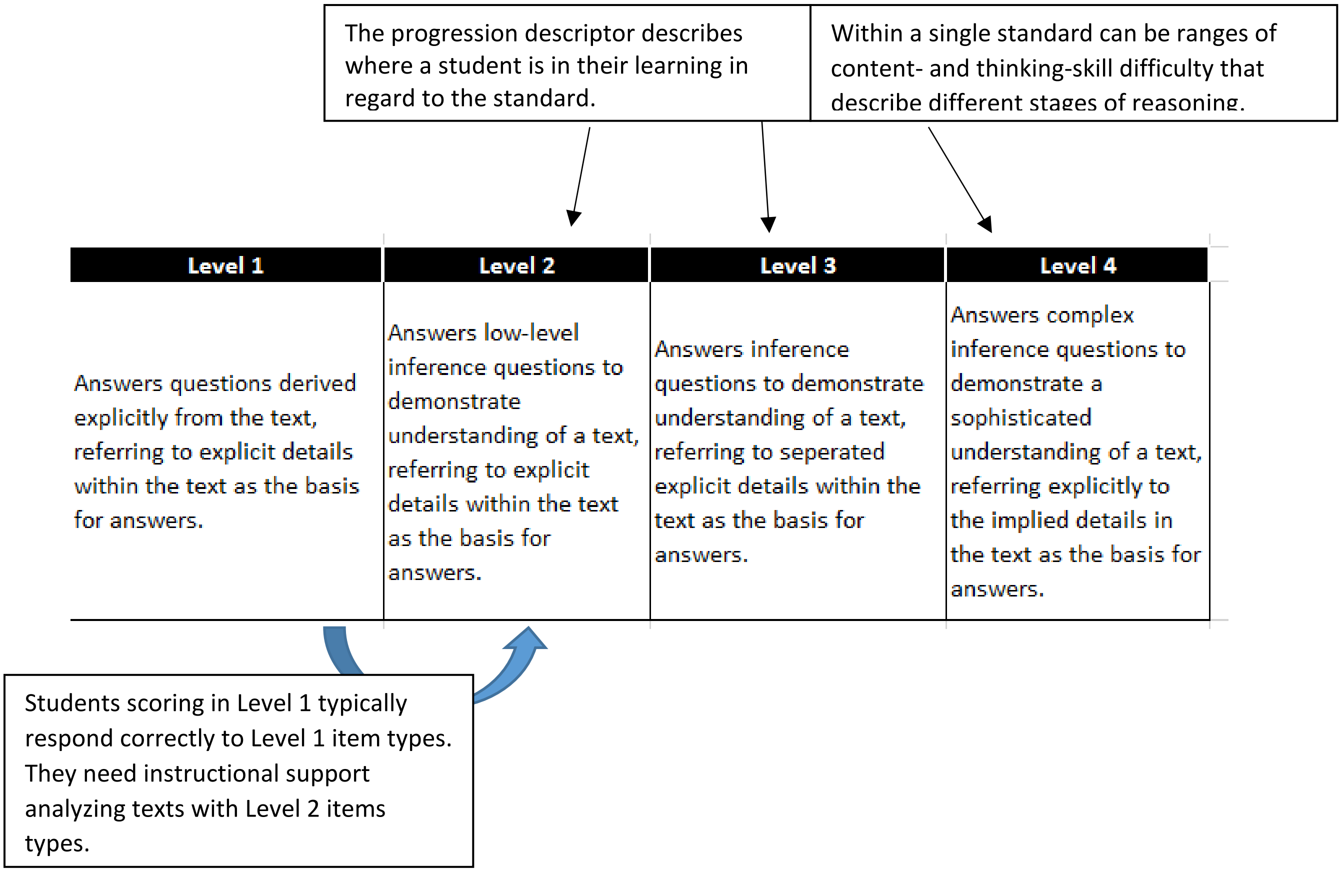
Within a single standard can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Level 1	Level 2	Level 3	Level 4
Answers questions derived explicitly from the text, referring to explicit details within the text as the basis for answers.	Answers low-level inference questions to demonstrate understanding of a text, referring to explicit details within the text as the basis for answers.	Answers inference questions to demonstrate understanding of a text, referring to seperated explicit details within the text as the basis for answers.	Answers complex inference questions to demonstrate a sophisticated understanding of a text, referring explicitly to the implied details in the text as the basis for answers.
DOK 1 EBSR	DOK 2 EBSR	DOK 3 EBSR	DOK 3 CR Multiple pieces of evidence

Figure 2: Range ALDS influence Item Specifications

Reporting ALDs – Reconciled Content Interpretation

Reporting ALDs are optimally created after final cut scores are adopted. More useful to teaching and learning is to examine the Range PLD trajectory and compare the hypothesized evidence of what students know and can do to the actual empirical evidence to determine if the trajectory has been confirmed or disconfirmed. Reporting ALDs need to reconcile what teachers and item writers hypothesized during the creation of the Range ALDs with the final cut scores that are influenced by the item difficulty, test use consequences, state policy, and the Range ALDs. Therefore, they should reflect the test construct based upon the final approved cut scores. The Reporting ALDs define the appropriate inferences stakeholders may make based upon the student's test score in relation to the final approved cut scores, and optimally teachers are given supportive information regarding how to interpret them to support formative practice as shown in Figure 3.



The progression descriptor describes where a student is in their learning in regard to the standard.

Within a single standard can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Level 1	Level 2	Level 3	Level 4
Answers questions derived explicitly from the text, referring to explicit details within the text as the basis for answers.	Answers low-level inference questions to demonstrate understanding of a text, referring to explicit details within the text as the basis for answers.	Answers inference questions to demonstrate understanding of a text, referring to seperated explicit details within the text as the basis for answers.	Answers complex inference questions to demonstrate a sophisticated understanding of a text, referring explicitly to the implied details in the text as the basis for answers.

Students scoring in Level 1 typically respond correctly to Level 1 item types. They need instructional support analyzing texts with Level 2 items types.

Figure 3: Supporting Teacher use of Reporting ALDS

Conclusions

The design and validation of an assessment system intended for both formative and summative purposes requires careful development processes, especially when such assessments are intended to support interpretations regarding how student learning grows more sophisticated over time (Pellegrino et al., 2016). Under a principled approach to assessment design, the evidence needed to draw a conclusion is made explicit in the ALDs and items are developed specifically to those evidence pieces (Huff et al., 2016; Egan et al., 2012; Schneider & Johnson, in press). Using a principled approach to assessment design is intended to support the validity of inferences about the stage of learning as well as the content validity of the assessment as a measure of student achievement over the course of a year. For this reason, states should develop item types to elicit authentic evidence of student development for each achievement level in regard to within-grade content. This direct connection to the achievement level descriptors will support teachers matching student work to the ALDs (Ferrara & Lewis, 2012), and based on the match, the next subsequent achievement level represents a likely instructional path appropriate for the student. This process moves the assessment to one that is a systemically valid assessment.

Teachers may more accurately locate a child along the “mastery of standards” construct in the future than they do in current practice by matching evidence from student work to the achievement level descriptors. A match is not a stopping point for decision making. It is a call to administer a more difficult task until a match cannot be made. It is a call to move deeper within the standards so that students have an opportunity to learn standards at a state’s intended levels of cognitive depth and content difficulty. The use of achievement level descriptors in this manner helps teachers interpret the student work evidence so that teachers can better identify where a child is in their learning and what the child needs next. Using a principles assessment design process, that concludes with test developers providing professional development and student work samples, supports teachers in better understanding a single standard has easier and more difficult representations (Figure 4) and that the goal of instruction is to support the development of student cognitive skills in addition to the content based skills.

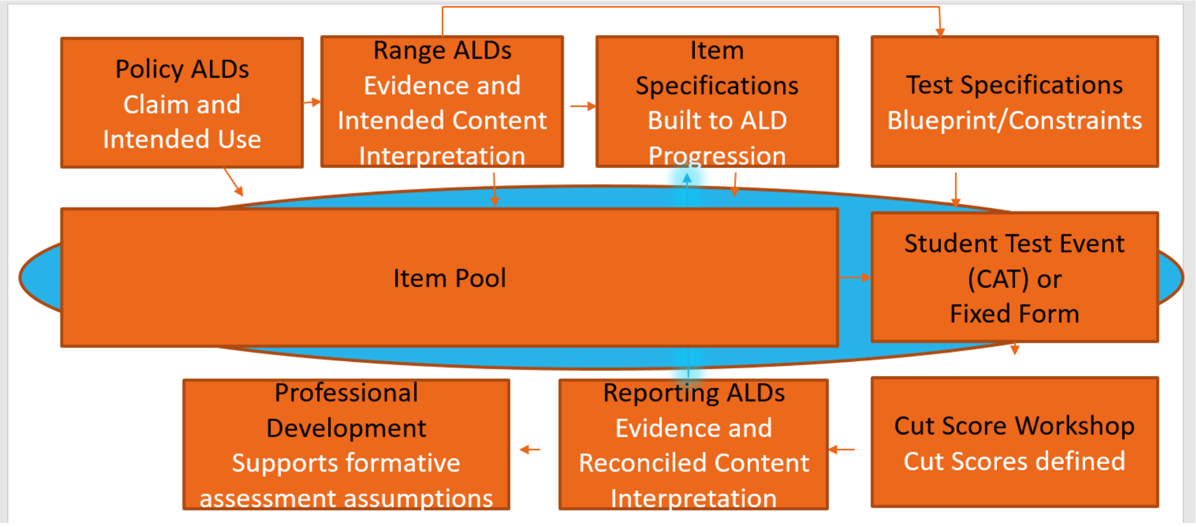
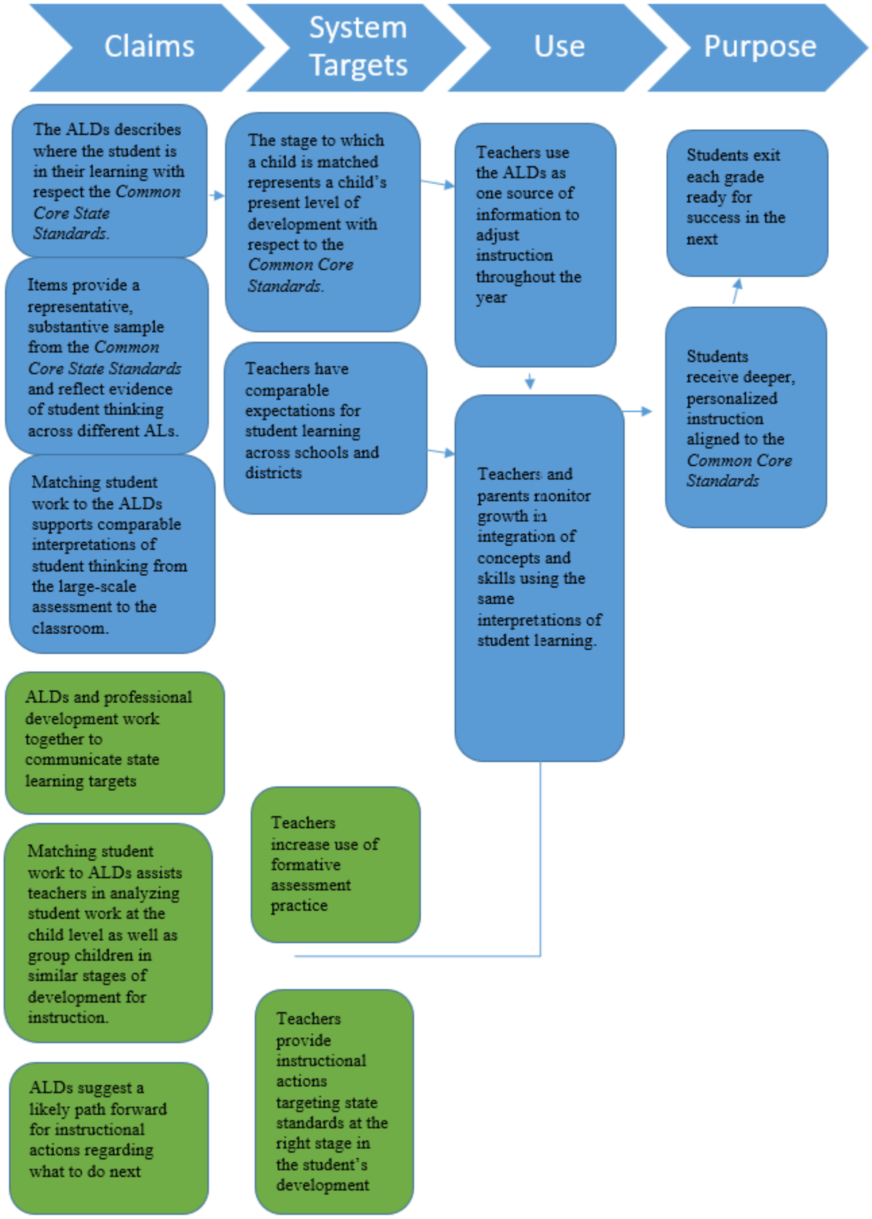


Figure 4: Principled Assessment Design to Support Systemically Valid Assessments

To achieve such a goal, achievement level descriptors and the assessment must be created using a principled design framework. They must be validated through research with a priori resolution rules regarding how to handle items that to do not match empirically their intended achievement level. Finally, ALD must be coupled with released samples of student work for teachers to analyze if a state desires to build an assessment system to support their primary intended test score use: informing instruction so achievement of students is accelerated. Figure 5 shows the theory of action for such an assessment system. As shown in Figure 5 the ultimate intended purpose of the system to have students exiting each grade ready for success in the next. The information generated from the assessments is intended as a source of information so teachers adjust instruction and monitor learning of individual students. The green text boxes represent the formative assessment assumptions that lay the foundation for the blue text boxes. Blue boxes represent the network of inferences and assumptions that will need to be validated across time, in addition to those typically associated with an assessment, to support the interpretive argument regarding what ALDs mean.



Claims

System Targets

Use

Purpose

The ALDs describes where the student is in their learning with respect the *Common Core State Standards*.

Items provide a representative, substantive sample from the *Common Core State Standards* and reflect evidence of student thinking across different ALs.

Matching student work to the ALDs supports comparable interpretations of student thinking from the large-scale assessment to the classroom.

The stage to which a child is matched represents a child's present level of development with respect to the *Common Core Standards*.

Teachers have comparable expectations for student learning across schools and districts

Teachers use the ALDs as one source of information to adjust instruction throughout the year

Teachers and parents monitor growth in integration of concepts and skills using the same interpretations of student learning.

Students exit each grade ready for success in the next

Students receive deeper, personalized instruction aligned to the *Common Core Standards*

ALDs and professional development work together to communicate state learning targets

Matching student work to ALDs assists teachers in analyzing student work at the child level as well as group children in similar stages of development for instruction.

ALDs suggest a likely path forward for instructional actions regarding what to do next

Teachers increase use of formative assessment practice

Teachers provide instructional actions targeting state standards at the right stage in the student's development

Figure 5: Theory of Action

References

Brookhart, S. M., Moss, C. M., & Long, B. A. (2008, March). *Professional development in formative assessment: Effects on teacher and student learning*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Brookhart, S. M. (2009). Editorial. *Educational Measurement: Issues and Practice, 28*(3), 1–3.

Carter, K. (1984). Do teachers understand the principles for writing tests? *Journal of Teacher Education, 35*, 57–60.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning, 6*(2), 81–89.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

Ferrara, S., & Lewis, D. (2012). The Item-Descriptor (ID) Matching method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 255–282). Routledge.

Ferrara, S., & Steedle, J. (2015). Predicting item parameters using regression trees: Analyzing existing item data to understand and improve item writing. *Paper presented at the annual meeting of the National Council of Measurement in Education*, Chicago, Ill.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Furtak, E. M., Morrison, D., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education, 98*(4), 640–673.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative classroom assessment? *Educational Measurement: Issues and Practice 28*(3), 24–31.

Hoover, N. R., & Abrams, L. M. (2013). Teachers’ instructional use of summative student assessment data. *Applied Measurement in Education, 26*(3), 219–231.

Huff, K, Warner, Z., & Schweid, J. (2016). Large-scale standards based assessments of educational achievement. In A. A. Rupp & J. P. Leighton, (Eds), *The handbook of cognition assessment: Frameworks, methodologies, and applications* (pp. 399–426).

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). Routledge.

Llosa, L. (2005). Assessing English learners’ language proficiency: A qualitative investigation of teachers’ interpretations of the California ELD standards. *The CATSOEL Journal, 17*(1), 7–18.

Marso, R. N., & Pigge, F. L. (1993). Teachers’ testing knowledge, skills, and practices. In S. L. Wise (Ed.), *Teacher training in measurement and assessment* skills (pp. 129–185). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

Moss, C. M., Brookhart, S. M., & Long, B. A. (2013). Administrators' roles in helping teachers use formative assessment information. *Applied Measurement in Education, 26*(3), 205–218.

Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice, 28*(3), 14–33.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist 51*(1), 59–81.

Perie, M., & Huff. K. (2016). Determining the content and cognitive demand for achievement tests. In S. Lane, M. Raymond, & T. Haladyna. *Handbook of Test Development (*2nd edition, pp. 119–143). Routledge.

Ruiz-Primo, M. A., Furtak, E. M., Ayala, C. C., Yin, Y., & Shavelson, R. J. (2010). Formative classroom assessment, motivation, and science learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative classroom assessment* (pp. 139–158). Routledge.

Schneider, M. C., & Johnson, R. L. (in press). *Creating and implementing student learning objectives to support student learning and teacher evaluation*. Under contract. Taylor and Francis.

Schneider, M. C., & Andrade, H. (2013). Teachers’ and administrators’ use of evidence of student learning to take action. *Applied Measurement in Education, 26*(3),159–162.

Schneider, M. C., & Gowan, P. (2013). Investigating teachers’ skills in interpreting evidence of student learning. *Applied Measurement in Education, 26*(3), 191–204.

Schneider, M. C., & Meyer, J. P. (2012). Investigating the efficacy of a professional development program in formative classroom assessment in middle school English language arts and mathematics. *Journal of Multidisciplinary Evaluation, 8*(17), 1–24.

Schneider, M. C., Smith, J., & Davidson, A. (2014, April). *Measuring teacher skill in formative assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Schneider, M. C., & Egan, K. (2014). *A handbook for creating range and target performance level descriptors*. The National Center for the Improvement of Educational Assessment.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 32–37.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1&2), 1–98.

Yap, C. C., Pearsall, T., Morgan, G., Wu, M., Maganda, F., Gilmore, J., . . . D'Amico, L. (2007). *Evaluation of a professional development program in classroom assessment*: 2006–07. University of South Carolina. Unpublished study.