

# Some Implications of the Design of Balanced Assessment Systems for the Evaluation of the Technical Quality of Assessments

Brian Gong  
Center for Assessment  
October 17, 2010<sup>1</sup>

## Introduction

Enhancing and documenting the quality of an assessment involves applying analytical tools to help ensure the assessment meets criteria for technical quality. Many criteria and guidelines for technical quality have been developed over the history of measurement. Corresponding conceptual and empirical tools have been developed to operationalize and help make more systematic the portrayal and evaluation of technical quality.

It has been envisioned by many that the next generation of “balanced assessment systems” will incorporate many new features that will advance what is typically done currently, particularly in large-scale assessment. For example, some envision these “next generation” assessment systems will more validly assess constructs not currently assessed well; will portray growth or progress of student learning; will be valid for a wider and more inclusive set of students; will provide information more directly useful for informing a larger set of efforts to improve by students, teachers, and school agencies that includes instructional action, program and personnel evaluation, and school accountability. In addition, many people hope the next generation of assessments will be more operationally flexible, more engaging, return results faster, provide more customized reports, and be less expensive.

If the next generation of assessments is to incorporate even some of these desired features, what implications do those changes have for our technical evaluation criteria and the ways we evaluate our assessments?

In this paper I explore this issue of the possible challenges of next generation assessments for technical quality criteria and tools by considering the topic of “growth.” An underlying theme is that we must define our constructs with more specificity and with different dimensions than we traditionally have done for large-scale assessments—and the differentiation of content and time

---

<sup>1</sup> The paper was prepared as background for a presentation/discussion session at the 2010 Reidy Interactive Lecture Series sponsored by the Center for Assessment and WestEd, held October 21-22, 2010 in Cambridge, MA. The paper is available at [www.nciea.org](http://www.nciea.org) under “Publications and Presentations.” I gratefully acknowledge the deep impact on my thinking about these topics from many conversations with my colleagues at the Center for Assessment, especially Charlie DePascale, Scott Marion, and Karin Hess. Much of the good here is due to them, and I accept responsibility for anything amiss that remains.

are particularly important. I consider three aspects of growth that require a greater specification of content and time:

- a) designing and interpreting a coordinated set of assessments that take place over time, such as interim assessments or “through course” assessments;
  - b) defining and measuring growth for individual students in relation to their curricular experiences, particularly for those students who are somewhat more advanced and those who are somewhat behind the intended curriculum; and
  - c) design and interpretation of the results from these multiple assessments.
- d) My fourth consideration is the tension this new attention to differentiated content and skills brings to our traditional scales as the means for interpretation, particularly for vertical developmental scales.

The paper concludes with stepping back from considerations of these four specific topics around growth to consider the implications that “next generation” assessment systems will have for our tools and criteria for technical quality of assessments.

### **A Note About the Examples**

This paper uses for its examples the *structure* of the construct and the corresponding *design* of assessments. It does not attempt to show the actual *structures of the disciplines* for particular content areas, although this has been a rich area of work, for example, by mathematicians, mathematics educators, and curriculum specialists. Similarly, the paper does not consider the psychological dimensions or *structure of tasks*, or *student cognitive models*, something that is of intense concern for item designers and cognitive scientists. And readers will note there is almost no discussion of the empirical results of tests—whether factor analysis supports the posited construct structure, for example. Attention in all of these areas is needed, together, to produce the next generation of assessments. This paper is a small invitation to begin that work with attention to defining the construct.

## Growth as Change in Performance Over Time<sup>1</sup>

- How do we define the assessment target, and how do considerations of content and time play into the specifications for assessment design and interpretation of results?
- How are the relations between multiple tests specified?
- What technical criteria and tools are available to help with this specification, measurement, and interpretation?

I'll first present the scenario, and end this section with some brief consideration of the technical quality tools and criteria.

### Scenario

The topic is portraying growth in terms of change in performance over time.

Assume that we want to measure growth in terms of change in student knowledge and skills over time. For this example, we present the discussion in terms of change within a school year, but we could extend it to a time period of less than or more than a school year. We present this first in terms of an interim assessment scenario. We will return to discussion of a “through course” design.

Interim assessments may be designed quite differently. Different designs will serve different purposes. Below are four test designs that differ in terms of the content included in a set of interim assessments administered four times during the year, followed by the state summative assessment.

Assume that during the school year the course of instruction includes ten topics, A-J, that are taught one per month, September through June. Some of the topics include subtopics, such as D1, D2, etc. This is a simplified version of a curricular/learning sequence to help make the points about the assessment design.

Simplified curriculum sequence of 10 topics/content standards during year										
Month	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
Topic	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	J

Assume that the student participates in five assessments: four interim assessments that are administered every seven weeks (October, December, February, and April), and that there is a

---

<sup>1</sup> A version of this section was developed for a project sponsored by CCSSO and generously funded by Renaissance Learning. The paper, “Balanced Assessment Systems” is under review by CCSSO.

single state assessment administered in June. Again, this is a simplified example of a testing schedule to illustrate the points.

Consider these four possible designs for the interim assessments in relation to the final state assessment.

***Interim assessment design 1: “The state test mirror” design***

Assume that the primary purpose for the interim assessments is to predict the student’s performance on the end-of-year state assessment. The theory of action is that if the student is predicted by the interim assessment to have problems passing the state assessment, then the teacher will use the information from the interim assessment to figure out what the student needs, help provide that, and the student will learn more and have a better chance of passing the state assessment in the end. As part of this, the theory is that the student needs exposure to the types of items on the state assessment and practice in performing in the state assessment context, and so the interim assessments will provide that practice, resulting in higher performance than if the practice had not been given in the interim assessments. A weaker version of this theory of action is that the teacher uses the interim assessment as a gross signaling assessment, but would use another assessment to diagnose what the student’s particular needs are.

<b>Simplified curriculum sequence of 10 topics/content standards during year</b>										
<b>Month</b>	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
<b>Content</b>	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	J

  

<b>Tests and Tested Content</b>					
<b>Test</b>	Interim #1	Interim #2	Interim #3	Interim #4	<b>State Test</b>
<b>Content Tested</b>	C, D4, F <sub>2</sub> , etc	C, D4, F <sub>2</sub> , etc.	C, D4, F <sub>2</sub> , etc.	C, D4, F <sub>2</sub> , etc.	C, D4, F <sub>2</sub> , etc.

In this design, the interim assessment mirrors the end-of-year state assessment in terms of content, balance of emphasis, format, administration conditions, etc. Each test administered during the year covers the same content and has the same design. The scores are as similar the state assessment’s as possible in terms of scale and achievement level to support the interpretation of how the student might do. Conversely, the performance on the interim assessment might be translated into a probabilistic projection, such as “The student has 75% probability of scoring at least at the Basic level”; this probabilistic projection might avoid reporting results of the interim assessment in terms of scale scores similar to the state assessment.

This design might provide high “practice” and high “prediction” from the interim to the end-of-year state assessment. It doesn’t provide much instructionally useful information for this year’s

teacher because the assessment topics are largely out-of-synch with the instructional sequence. For example, many topics are tested before they are taught, e.g., the complete first interim assessment consists of topics that have not yet been covered by October. (Interim assessment #1 covers content C, D4, F2, etc., when the student has been instructed on content A and B.) Many people would argue that this design of interim assessment provides very little information that is instructionally useful, and that if the teacher needs this interim assessment to know that the students do not yet know the content, then there is probably something wrong with the teacher’s ability to assess. On the other hand, if there are students who do know the content for these topics before the instruction occurs, then one might question why they are enrolled in this class, and certainly might foresee problems with the planned curriculum sequence.

Technical tools and criteria: Many people would say it would be easy to place the results of the interim assessments on the same scale with the other interim assessments and with the state assessment. Similarly, many people would say that it would be easy to evaluate the quality of alignment of the interim assessments to their learning targets because it is essentially the same task as evaluating the alignment of the state test to its learning targets; there are some recognized methodologies and criteria for such alignment analyses.

***Interim assessment design 2: “The non-cumulative instructional mirror” design***

Assume that the primary purpose for the interim assessments is to help inform the teacher about what the student has mastered or not after instruction so that the teacher can take appropriate actions. The theory of action is that if the teacher knows what the student did not learn well enough, the teacher can remediate before the deficits get too large. An assumption is that if the student knows the material well at the time instructed, then the student will be able to perform satisfactorily on the state assessment at the end of the year.

Simplified curriculum sequence of 10 topics/content standards during year										
Month	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
Content	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	J
<b>Tests and Tested Content</b>										
Test	Interim #1		Interim #2		Interim #3		Interim #4		State Test	
Content Tested	A <sub>12345</sub> , B <sub>1234</sub> ,		C <sub>123456</sub> , D etc.		E etc., F etc.		G etc., H etc.		C, D4, F2, etc.	

In this design, the interim assessment focuses on the content that was instructed. Each interim assessment covers only the content in the most recent instructional period, and thus each test’s content differs from the others. The assumption is that this interim assessment is closely connected with an effect instructional program that takes the information from the interim assessments about student needs and informs effect instructional supports. The efficacy of the

instructional interventions will mediate the relationships of student performance scores on the interim assessments and the state test (and perhaps of scores between the interim assessments).

The interim assessment should provide more detail about student performance on subtopics. The selection of subtopics is very important if one is interested in supporting future student learning and performance: one might focus on topics which are most important for facilitating the subsequent learning, or one might focus on assessment in a “mastery learning” mode that reflects what was instructed.

***Interim assessment design 3: “The total cumulative instructional mirror” design***

Assume that the primary purpose for the interim assessments is to inform the teacher and student of the student’s mastery of all the course content. The state test is viewed as an incomplete sampling because of constraints on time and perhaps different priorities. The theory of action is that cumulative learning is what is desired—either because it is felt all the knowledge and skills are important for the student to remember, or because it is felt that proficiency consists of the student integrating the various knowledge and skills over the course. Since cumulative learning is what is desired, the interim assessments should reflect that and inform it.

<b>Simplified curriculum sequence of 10 topics/content standards during year</b>										
<b>Month</b>	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
<b>Content</b>	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	J
	<b>Tests and Tested Content</b>									
<b>Test</b>	Interim #1		Interim #2		Interim #3		Interim #4		State Test	
<b>Content Tested</b>	A, B		A, B, C, D		A, B, C, D, E, F		A, B, C, D, E, F, G, H		C, D <sub>4</sub> , F <sub>2</sub> , etc.	

In this design, the interim assessment is designed to assesses what was instructed, but is cumulative, i.e., the assessment includes all topics instructed up to that point in time. The nature of the topics and their relations might change over time, so that some content knowledge and skills taught and assessed earlier in the year is assessed for different properties later, such greater expertise as reflected in greater automaticity, greater connections and integration, more application or extensions, or more judgment about the limitations of such content or contextual awareness of when to apply it.

***Interim assessment design 4: “The non-cumulative instructional mirror” plus cumulative state assessment mirror” design***

Assume that the primary purpose for the interim assessment is to help inform the teacher about the student’s mastery of recently instructed content, but also whether the student is maintaining performance on that content and skills that is likely to be assessed on the state assessment.

Simplified curriculum sequence of 10 topics/content standards during year										
Month	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
Content	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	J
Tests and Tested Content										
Test	Interim #1		Interim #2		Interim #3		Interim #4		State Test	
Content Tested	A, B		B, C, D		C, D, E, F		C, D <sub>4</sub> , F <sub>2</sub> , G, H		C, D <sub>4</sub> , F <sub>2</sub> , etc.	

In this design, the interim assessment is designed to assess what was recently instructed, but is also cumulative for the topics that will be assessed on the state assessment.

### Discussion of Implications for Technical Criteria and Tools

The four interim assessment designs could all be used to predict how well a student might do on the state assessment, for example. However, the designs have different assumptions (notably about students’ remembering/forgetting over time) and very different score structures. For example, a score of “50% of the items correct” would have very different interpretations of how well a student was prepared, for the various designs, where 50% would be a high performance in the first design (since the student has not yet been instructed on most of the content included on the test), and a low performance in the second design (where ostensibly the content was all taught recently).

Note that in every one of these designs, 100% of the test items are aligned to the state’s content standards (one-way alignment between items and standards). Designs 2 and 3 would probably fail a typical alignment study using the typical criteria (e.g., Webb’s range of representation) if the interim assessments were evaluated for degree of alignment to the complete state standards, the test specifications for the state test, or the state test form/items. Design 4 might also be judged to have “weak” alignment, depending on the overlap between the final state test and the interim assessments and instructional content targets.

Do we think these content distinctions are important? Can our scales reflect these differences?

### Some Other Important Applications

This type of analysis can be applied to other important assessment topics. I briefly introduce three such applications:

- A “look ahead” model of instruction that changes the timing of what is included on an interim assessment
- The design of end-of-course exams, and the possible relations of end-of-course exams to an end-of-grade-span exam
- The design of “through course” assessments

***Interim assessment design 5: “The post-/pre-assessment” design***

The “instructional mirror” design 1 has an implicit theory of action that makes instruction largely reactive to assessment. That “goal-gap-reduction” theory of action casts instruction and learning in a model of reducing the gap between a desired goal state and the current state. The assessment provides feedback about the effects of an action so that subsequent action can be modified. In this model, feedback is always looking backwards at what has been done, and planful action is not modified until after feedback is received. In terms of instruction and assessment, in this model instruction-informed-by-assessment is always remedial because the assessment only informs about instruction that has already occurred.

A contrasting model seeks to inform instruction before it happens.

<b>Simplified curriculum sequence of 10 topics/content standards during year</b>										
<b>Month</b>	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
<b>Content</b>	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	J
<b>Tests and Tested Content</b>										
<b>Test</b>	Interim #1		Interim #2		Interim #3		Interim #4		State Test	
<b>Content Tested</b>	A, B, C, D		C, D, E, F		E, F, G, H		G, H, I, J		C, D <sub>4</sub> , F <sub>2</sub> , etc.	

In this design, the interim assessment is designed to assess what was instructed, but also to provide a preview to the teacher of student knowledge about the topics coming up in the next instructional time period. The theory of action here is that teachers can help students by knowing their strengths and weaknesses prior to instruction, so they can change instruction to fit where the students are. This theory of action may be contrasted with one that focuses on maximizing student learning by instructing-assessing to see where the problems are-and then providing remedial instruction. In the latter view, the adaptive instruction always occurs after the main instruction; in the former view, adaptive instruction occurs both as part of the main instruction and after.

There is much discussion and debate about what the student knowledge-and-skill models are that might be profitably assessed (e.g., what grain-size, how stable they are, and how idiosyncratic they are), what the nature of pre-assessments might be that would allow informative assessment, and what the schooling structures might be that would enable such look-forward-adaptive instruction.

***End-Of-Course assessment design 1: “What is cumulative?”***

These same analyses can be extended to end-of-course (EOC) assessment, where the course and EOC exam are the units of analysis rather than interim assessments within a year’s course.

<b>Simplified curriculum sequence of 9 topics/content standards over three courses</b>										
<b>Course</b>	Course I			Course II			Course III			End of Grade 11
<b>Content</b>	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I	A-I
<b>Test</b>	EOC #1			EOC #2			EOC #3			<b>Grade Span Test</b>
<b>Content Tested</b>	A <sub>12345</sub> , B <sub>1234</sub> , C <sub>123456</sub>			D <sub>145</sub> , E <sub>2356</sub> , F <sub>12367</sub>			G <sub>4</sub> , H <sub>2357</sub> , I <sub>1234</sub>			B <sub>1</sub> , C <sub>2</sub> , D <sub>4</sub> , F <sub>2</sub> , etc.

EOCs are almost always considered to have a summative evaluation function, so some of the discussion is whether an EOC can also have an instructional function for the students who take the EOC that year. Assuming that the EOC is about an hour in length, the combination of EOCs could be much longer and assess the content in more detail than the typical state end-of-grade-span test.

An important question is how a grade-span assessment might be designed to assess something valuable and different from the combination of end-of-course exams. For example, the end-of-grade-span test might be designed to assess what students could do at that point in time (say, end of grade 11), rather than assuming that a student who passed the EOC #1 in grade 9 still remembers the content and skills two years later. Another possible example is an end-of-span test might test for integration and application of the content knowledge and skills that were learned in separate courses and tested separately (e.g., the ability to apply what was learned in the Algebra I course and in the Geometry course to solve problems that involve knowledge and skills from both courses).

***“Through Course” assessment design 1: “What counts as cumulative?”***

A “through course assessment” design has been proposed for use by at least one of the federally funded common assessment consortia. There are many possible through-course designs; I am not aware of a detailed specification that has been adopted by any of the main groups that use the term, “through course assessment.”

- One possible aspect of a through-course design would be to have the test at the end of the year non-cumulative, i.e., certain content and skills would only be assessed by certain through-course assessments that take place earlier in the year.

- A possible aspect of through-course design would be to have the through-course assessment earlier in the year represent complex knowledge, skills, and performances that would continue to be taught after the performance were assessed. For example, a research paper and class presentation would constitute the through-course assessment in English and would be done by the student in February (primarily to allow time for human scorers to score the assessments and return the results by the end of the school year).
- A third possible aspect of through-course design would be to have certain aspects of content assessed both by the more complex through-course assessments (presumably some type of constructed response or performance assessment) and also assessed through the end-of-year assessment consisting entirely of machine-scorable items (presumably multiple choice and cousins).

These three possibilities are portrayed below in terms of the different content tested.

Simplified curriculum sequence of 10 topics/content standards during year										
Month	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
Content	A	B	C	D <sub>1234</sub>	E	F <sub>123</sub>	G	H	I'	J

  

Test	Through Course #1	Through Course #2	Through Course #3	Final Test
Content Tested	A <sub>12345</sub> , B <sub>1234</sub> , C <sub>123456</sub>	D <sub>145</sub> , E <sub>2356</sub> , F <sub>12367</sub>	G <sub>4</sub> , H <sub>2357</sub> , I <sub>1234</sub>	J
Content Tested	B, E, H, I	C, F, H, I	G, H, I'	B <sub>1</sub> , C <sub>2</sub> , D <sub>4</sub> , F <sub>2</sub> , etc.
Content Tested	B <sub>1</sub> , C <sub>2</sub> ,	D <sub>4</sub> , F <sub>2</sub>	G <sub>4</sub> , H <sub>2</sub>	B <sub>1</sub> , C <sub>2</sub> , D <sub>4</sub> , F <sub>2</sub> , etc.

It is possible to think of other desirable content schemes, such as a “spiral curriculum” where certain concepts or skills appear more than once in the course of instruction.

Scenarios such as the second above, where Content/Skill I is tested before it is taught are very reasonable when “I’” is a variant of something that could be acquired by the student previously, such a writing skills. It would be quite possible to ask a student to write an analysis of a literary work at the beginning of the year, knowing that the curriculum will teach the student more advanced skills in analysis later in the year; the table above demarcates I’ from the previously learned I. But this notion of having to specify and perhaps assess previous related knowledge makes the enterprise more complex.

This type of consideration of content sequencing makes considerations of measuring “value-added growth” more complex.

## **Growth Through Individualizing the Implemented Curriculum**

The first section in the paper looked at several ways to define periodic assessments (interim, pre-assessment, end-of-course, and through-course) in relation to a hypothetical course of instruction and a state test given toward the end of instruction. This section discusses the need to define growth when the student grows not in concert with the generally defined curriculum.

Many states are interested in measuring “student growth,” where growth is defined as the change in performance. This section makes two fundamental points about measuring student growth:

1. The type of learning must be defined in order to validly measure growth; and
2. Learning is usually productively defined by instruction, not just content standards.

Of course, the more care we take about our measurements and comparisons, the better our descriptions will be.

### **Scenario: Defining Desired Learning More Advanced Than the Intended Curriculum**

Envision the case of a student who has learned the intended curriculum’s content and skills by Thursday, and the teacher has planned to spend until Friday on this unit. What could the teacher ask the student to do? If a student were proficient at the first point in time, what should s/he learn in class next? How would that define our definition and measurement of growth? (Although I’ve framed this scenario in terms of student progression within a week, the same issues hold for other units of learning, such as units and courses.)

Four possibilities include:

1. “Super Mastery”: The student could go back and study to be sure s/he knew everything taught, 100%. If the performance were portrayed as a test score, the student would study whatever s/he missed on the test so that s/he could try to get a perfect score.
2. “Curricular Advancement”: The student could take what had been learned and go on to the next topic in the curriculum sequence. If the performance were portrayed as a test score, the student would prepare for the next unit’s test that would cover material in the next unit of instruction.
3. “Expertise/Application Development”: The teacher might challenge the student to learn to apply what had been learned in new or more complex settings, or do something independent with it, or in other ways develop deeper expertise with the content that had been learned in that unit.
4. “Student Choice”: The teacher might give the student permission to learn something unrelated to the topic, or whose relationship was unspecified by the teacher. For example, the teacher might say, “Why don’t you read a book you like?”

Assume that the student actually does learn something additional. To be able to assess what that growth was, we would need to have an assessment that was sensitive to the type of learning: “Super Mastery,” “Curricular Advancement,” “Expertise/Application Development,” or “Student Choice.”

The type of desired learning must be specified, and the test be developed to assess that type of desired learning. Otherwise the measurement of student growth will not reflect accurately either what the student knows and can do at the second point in time, nor will it reflect the student’s growth over time along the construct or dimension of interest.

However, because my supposition was that all these examples of growth, except the first of “Super Mastery,” go beyond the specified curriculum, it implies that measurement of this type of growth will need to change the model of standards-based assessment currently widely used. In that model, assessment items are to be tightly aligned to the content that is to be learned, presumably reflected in the well-aligned curriculum. Alignment criteria would mark “Curricular Advancement” as out of grade level, while “Expertise/Application Development” would show up as being a depth-of-knowledge too high and/or content/skills not included in the content standard or in the achievement level descriptors.

### **Scenario: Defining Growth Not as Advanced As the Intended Curriculum**

A similar scenario can be conceived to define and assess “growth” for students whose learning involves “catching up” to the intended curriculum.

1. “Mastery of Parts”: The student is expected to learn some but not all of the same content included in the intended curriculum, to the same level of mastery as expected in the intended curriculum.
2. “Mastery of Prerequisites”: The student is expected to learn the prerequisite or precursor knowledge of skills defined in the intended curriculum before this unit or grade.
3. “Different Degree of Expertise”: The student is expected to learn the same content but not at the same level of expertise. For example, the student may require more scaffolding or other assistance or the student may not do as much generalization to new contexts or problems. Demonstrating the levels of expertise may require different instructional or assessment contexts than typically provided by the intended curriculum or regular assessment.

Similar to the growth more advanced than the intended curriculum, to capture the growth achieved by the student, the assessment must be sensitive to the dimensions of possible learning, and reflect the goals and instructional strategies.

## **Interpreting Growth**

If one takes an existing test or set of tests, and uses the test scores to calculate growth without specifying what the desired growth should be and ensuring the test does measure that desired learning appropriately, then one may get a number, but the growth number is not necessarily interpretable. This is particularly true for the conditions of growth at the higher and lower ranges of performance because the learning may be taking place in different dimensions than specified by the intended curriculum.

For the condition of “Expertise/Application Development” in particular, it is necessary to know what the instructional history has been for the student. This is because one cannot tell if a particular performance is really an application without knowing whether the student had been directly instructed on it previously.

## **Recommendations**

1. Specify the desired learning in sufficient detail that you and others can tell when it is occurring. Attend to how much is “super-mastery,” “application,” and “advancement.”
2. Ensure the assessment measures the learning you desire, especially over time.
3. Ask for content-based interpretations of assessment scores—especially assessment scores such as “vertical scales” that claim to portray growth—to allow you to check whether the “high scores” produced by the assessment correspond to your definition of desired learning.
4. A test must be carefully constructed in order to be used to assess growth. For example, a test assembled from an item bank must refer to other information, such as definition of desired learning and instructional history, to be used to assess student growth. The typical statistical information associated with test items (e.g., difficulty, coding to general content standard, lack of bias) is insufficient to adequately specify whether an item is appropriate to measure growth.

A challenge to almost all modern sets of content standards is that they are very thin on depictions of expertise, performance, or quality in general. I think this stems in large part from the initial split of “content standards” from “performance standards.” When content and performance standards were closely attached to tasks, then there was more rich and detailed depiction of quality. In general, the predominant focus of the standards movement on elaborating content standards has not been accompanied by similar development of how to portray expertise (other than lists of content). The two main places expertise is typically portrayed in current standards-based assessment systems are a) in the verbs of content standards and b) in achievement level descriptors. Both are too thin to provide real distinctions or guidance on designing assessments that differ in levels or kinds of expertise assessed.

Getting depictions of quality of understanding and performance that can be linked to performance standards would be a major contribution, in my opinion. What constitutes good understanding of crucial concepts? What defines good performance of critical types of tasks?

If the set of criteria are only applied to content standards, divorced from performance quality, then we will continue to have the same problems that we have now, of which a chief problem is that the content and performance standards do not carry enough critical information and so people look to assessment specifications and the tests items themselves to tell them what “readiness” or proficiency really means or what the learning targets really are.

## Growth and Multiple Measures

The prospect of a system of multiple assessments raises the issue of how to combine the results from multiple measures to inform a coherent portrayal or judgment of growth.

One essential design question about multiple measures is whether the multiple measures are intended to provide a broader, more complete assessment of the construct of interest, or whether the multiple measures are intended to provide triangulation information about the same construct of interest. To use concepts from validity, the first design is concerned with reducing construct under-representation, while the second design is primarily concerned with detecting or avoiding construct-irrelevant variance. It is possible, of course, to do both, but it is essential that the design reflect the intention. In terms of evidence-centered design, the design of the assessment should reflect the type of evidence required to inform the inferences and actions to be made.

<b>Design Considerations for Multiple Measures in Terms of Validity Concerns</b>				
Concern	Design	Example Assessment Target	Measure 1	Measure 2
Reduce error due to construct-under-representation when each assessment measure is known to incompletely assess the whole construct	Complementary content (or skills, performance levels, etc.)	AB	A	B
Detect/reduce error due to construct-irrelevant variance when occasion or person is thought to be a factor	Repeated administration of the same assessment	A	A – occasion 1	A – occasion 2
Reduce error due to construct-under-representation and construct-irrelevant variance when each assessment is known to incompletely assess a part of the construct	Variations in assessing the same construct	A	A'	A''

The assessment design for multiple measures should reflect the validity concern for the set of assessments. Stated positively, the set of assessments should be designed to present sufficient evidence to inform the intended interpretations and uses/actions, especially to appropriately minimize construct-underrepresentation, construct-irrelevant variance, and unintended negative consequences.

The function of the multiple assessment measures is informed by the role they play in terms of enhancing validity. The design of the assessment items is influenced by their relationship as a whole set. And the scoring and interpretation will naturally need to reflect this design.

I would like to discuss performance assessments within this framework of addressing validity concerns. Each performance assessments—like any other assessment item, task, or test—should be designed explicitly to provide at least one of the types of evidence above. Is the performance assessment adding to the validity by measuring something quite different that is not measured by the other assessment components? Is the performance assessment measuring the same construct but in a different way than the other assessment components? Is the performance assessment measuring the same construct in a way that mode doesn't matter but is providing more assessment occasions of measurement?

The approach I will take is to discuss performance assessments in terms of the rationales provided for using them. The same discussion could be extended to “innovative assessments” as they are being termed in some current assessment documents.

### **Performance Assessment: Conceptions and Rationales<sup>1</sup>**

Over the past several years there have been three major conceptions of performance assessment, each associated with different rationales for its nature and importance. The three conceptions have been focused on assessment format, the nature of learning, and the role of assessment in changing the nature of schooling.

1. **Format:** This conception of performance assessment involves students constructing responses rather than selecting responses.

The focus is on the *format* of the performance.

Rationales include:

- A. Assessment formats beyond multiple choice can provide evidence about student thinking not available in multiple choice exams (e.g., “showing approaches to solving a math problem”),
- B. Performance assessment formats and scoring provides evidence about complex answers (e.g., allowing partial credit),
- C. Assessment formats involving student construction of responses invokes different student thinking than does student selection from multiple choice responses (e.g., the erroneous perception that all multiple choice questions tap into ‘lower-order thinking’ and all constructed response questions tap into ‘higher-order thinking’ or the plausible perception that the thinking strategies involved in deciding among four given choices are different from those involved in generating a response).

---

<sup>1</sup> This subsection is based on an unpublished paper I wrote for a technical advisory committee meeting on assessment and accountability convened by Achieve.

- D. The hope (little supported by studies<sup>1</sup>) that different modes of responding would reveal that students who scored low on multiple-choice format tests would be able to demonstrate their true, higher ability in different modes.
- E. Teaching, particularly preparing students for a test, would be different and better if instruction and test-prep included more than multiple-choice formats. (See rationale 3.)

**2. Nature of Learning:** This conception of performance assessment stems from a concern that students learn more complex content and skills which can only be demonstrated well through performances.

The focus is on the *nature of learning*, both *content and process*.

Rationales include:

- A. The learning targets/content are inherently complex, abstract, or involve application or generalization such that student performances are the only way to assess validly. For an example, see the descriptions of authentic mathematical activities, the nature of real mathematics (mathematical models), deeper and higher-order understandings of a conceptual model, and deep and higher-order understandings in elementary mathematics by Lesh and Lamon, especially their lists of characteristics on pp. 18, 26, 30-31, and 32-33. Similar arguments have been made in other content areas (e.g., by Duschl (1990) in science, Baker (1990) in history, and Wiggins (1993) in English language arts. A similar argument, but about very different content and skills, has been made for “habits of mind” for higher education and work; see B below.
- B. The desired competencies involve more than academic knowledge and skills. Strategic and metacognitive problem-solving skills, “real world common sense,” problem solving skills, and ability to apply academic knowledge in common contexts are needed. In addition, there are many “soft skills” (e.g., interpersonal communication and problem solving) that are needed (e.g., how to get and use feedback appropriately within a writing process; how to work together on a team when the problem is such that one person alone could not solve it well). Many of these skills are not commonly taught in school because they are specialized or not valued. (See Conley, 2007.)
- C. Education ultimately is about students being able to produce and to continue to learn on their own. Assessments should be designed and support the progressive development of student ability and responsibility for their own learning. For example, Wiggins (1993) argues for performance assessment within a larger framework of a liberal arts tradition. He presents eight dilemmas and nine postulates about assessment that reflect this view of the goals of education and the nature of learning. (See Appendix.)
- D. Assessments should invoke the knowledge and skills and provide direct evidence, not merely correlational or predictive evidence. That is, there is assessment value in having students be assessed on tasks directly centered on the construct of interest, even though

---

<sup>1</sup> But most studies of “performance assessment” in education have been quite limited, using only written constructed response in formal, standardized settings as the only alternative to multiple choice.

performance on those tasks may be highly related to other scores or data. For example, writing scores on an essay prompt may be highly correlated with scores of vocabulary on a multiple choice exam, and ability to learn to run certain machinery may be highly related to a test of digit-span-memory and mental rotation of two-dimensional representations of three-dimensional shapes, or scores on a state math assessment may be correlated .90 with SES, but some people would like assessment evidence from the task or construct itself of writing an essay, operating the machinery, or solving math problems.

- E. The desired competencies are not commonly assessed because of some or all of the reasons above, and assessment is a way to bring attention and get the system to help students learn new things.
- F. Performance assessment can be more motivational to students because they involve more relevant settings and tasks, and higher student engagement. In many cases students can engage in meaningful performances (e.g., job apprenticeships) or design their own performances (e.g., projects) that are recognized as valuable evidence, whereas most students could not (nor would wish to) construct a traditional “test.”

**3. Leverage for Changing the Nature of Schooling:** This conception of performance assessment stems from a concern that students should learn things that are not usually available in schools due to the nature of school organization and values; those desired things may include cross-disciplinary integration of knowledge and skills, the ability to apply what is learned in non-trivial ways, and student-directed learning (including self-evaluation and pursuit of areas of interest to the student).<sup>1</sup>

The focus is on student performance as a powerful lens for shaping the nature of school organization and student interactions within school.

Rationales include:

- A. There is a large divide between schools and life outside of schools in terms of what is important to learn, how to learn it, and how to use that knowledge and skills. School should be reshaped to better prepare and engage students in “real world” learning and performance.
- B. Educated citizens and workers must integrate knowledge and skills, but school presents them in academic “silos.” Complex, integrated performances are a way to push student learning and school instruction towards what is ultimately needed.
- C. Examination of student work is one of the most powerful ways to get teachers and administrators to view and improve their work. Student work that is performance-based

---

<sup>1</sup> A very influential expression of this viewpoint was Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O’Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers. This also became the conceptual framework for the influential work of the *New Standards Project*, which purposefully incorporated extensive constructed response format items in its *New Standards Reference Exam*.

allows more insight into what teachers and administrators can and should do to improve teaching and learning than do other forms of assessment evidence that are more remote, such as multiple choice exams, formulaic worksheets, and abstracted summaries such as grades.

- D. Assessment is the most powerful and cost-effective way to change the educational system. See, for example, the arguments by Joe McDonald et al. (1993) for graduation by exhibition “assessing genuine achievement”), set within the context of the efforts of the Coalition of Essential Schools. McDonald starts his chapter, “This book is about a strategy for school reform called *planning backwards from exhibitions*.” (p. 1).

The specification of “what is (a) performance-based assessment” depends upon the conceptualization. A performance assessment designed to reflect a certain rationale will have certain characteristics, while other rationales will call forth other characteristics. The simplistic attention to surface features of performance assessments or listing of types—including portfolios, projects, reports, performances, problem solutions, and simulations—must be supplemented by a specification of what the construct is and valid ways to assess it. This paper suggests that identifying the key rationales is one fruitful way to begin specifying the construct.

There was high interest in performance assessment in the 1990s. Tucker (1991) gives a short historical perspective that emphasizes the political consensus that evolved through the 1980’s and resulted in assessment being “issue Number One.” Tucker, reviewing the very rapid change starting with *A Nation At Risk* (1983), boldly stated, “There is, I think, some mysterious unseen hand that shapes the consensus in American education. Almost overnight, assessment has become the focal point of a great debate about the purpose, shape, and control of American education, and it seems more likely than not that, within a decade, the United States will have a national examination system.” (p.3)

The hopefulness of the movement, as well as recognition of emerging challenges is reflected in the book edited by Baron and Wolf (1996), *Performance-based student assessment: Challenges and possibilities*. The editors’ preface provides an excellent overview of the landscape and a now somewhat poignant assessment of what it might take for performance assessments to be more widely used. The book’s organization reflects continuing issues: “Toward access, capacity, psychometric soundness and coherence,” “Realizations at the district and state levels,” and “Possibilities at the national level.”

The volume edited by Kulm and Malcom (1991) offers a different view of the key factors in educational reform in its section headings: “Policy issues in science assessment,” “Science assessment and curriculum reform,” and “Science assessment in service of instruction.”

Baker (2009) provides an update on the research base for performance assessments and also provides an appraisal of the current socio-political context in which assessments play, including

the ascendancy of accountability and more emphasis on the technical qualities of large-scale assessments than on their integration with instruction.

The National Academy of Science's Board on Testing and Assessment sponsored in 2009 a workshop that asked what lessons could be learned from past efforts of performance assessment with the hope of informing, "How can the myriad problems with performance assessments be solved: technical, cost, political, etc.?"

#### References

- Baker, E. L. (1990). "Developing comprehensive assessments of higher order thinking." In G. Kulm, (Ed.), *Assessing higher order thinking in mathematics*. Washington, DC: American Association for the Advancement of Science, pp. 7-20.
- Baker, E. L. (2009). The influence of learning research on the design and use of assessment. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 333-355). New York: Cambridge University Press
- Baron, J. B. & Wolf, D. P. (1996). "Editorial Preface." In J. B. Baron & D. P. Wolf, (Eds.), *Performance-based student assessment: Challenges and possibilities*. Ninety-fifth Yearbook of the National Society for the Study of Education (Part I). Chicago, IL: University of Chicago Press.
- Cohen, M. (13 Aug. 2007). Letter from Achieve to House Committee on Education and Labor, with recommendations regarding performance assessment
- Conley, D. (2007). *Toward a more comprehensive conception of college readiness*. Eugene, OR: Education Policy Improvement Center.
- Duschl, R. A. (1990). *Restructuring science education: The importance of theories and their development*. New York: Teachers College Press.
- Kulm, G. (Ed.). (1990). *Assessing higher order thinking in mathematics*. Washington, DC: American Association for the Advancement of Science.
- Kulm, G. & Malcom, S. M. (Eds.). (1991). *Science assessment in the service of reform*. Washington, DC: American Association for the Advancement of Science.
- Lesh, R. & Lamon, S. J. (1992). "Assessing authentic mathematical performance." In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics*. Washington, DC: American Association for the Advancement of Science.
- McDonald, J. P., Smith, S., Turner, D., Finney, M., & Barton, E. (1993). *Graduation by exhibition: Assessing genuine achievement*. Alexandria, VA: ASCD.
- Tucker, M. S. (1991). Why assessment in now Issue Number One. In G. Kulm & S. M. Malcom (Eds.), *Science assessment in the service of reform*. Washington, DC: American Association for the Advancement of Science, pp. 3-16.
- Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass.

## Growth in Content and Vertical Scales

It would be very desirable to be able to measure growth along a single scale. Some states have a vertical developmental scale that spans multiple assessments from multiple grades. Some commercial assessments also have vertical scales.

The issue is how to interpret progression along a vertical scale. What does it mean to have the same sequence of scale scores representing student progression, when those scores are from different tests?

I'll provide an example to inform this discussion.

Assume the table below provides the scale scores for five reading tests that are aligned with the content standards in: grades 3, 4, 5, 6, and 7. The tests share a common vertical scale ranging from (at least) 86 through 2767. (86 is the lowest reported scale score on the grade 3 test, and 2767 is the highest scale score reported on the grade 7 test.) Further, five achievement levels have been established for each test, called Level 1, Level 2, and so on up through Level 5. The cutscores for each achievement level have been established as shown in the table. For example, for grade 3, Level 2 ranges from 1046-1197.

<b>Grade</b>	<b>Level 1</b>	<b>Level 2</b>	<b>Level 3</b>	<b>Level 4</b>	<b>Level 5</b>
<b>3</b>	86 - 1045	1046 - 1197	1198 - 1488	1489 - 1865	1866 - 2514
<b>4</b>	295 - 1314	1315 - 1455	1456 - 1689	1690 - 1964	1965 - 2638
<b>5</b>	474 - 1341	1342 - 1509	1510 - 1761	1762 - 2058	2059 - 2713
<b>6</b>	539 - 1449	1450 - 1621	1622 - 1859	1860 - 2125	2126 - 2758
<b>7</b>	671 - 1541	1542 - 1714	1715 - 1944	1945 - 2180	2181 - 2767

1. Let's say Student A is in grade 4, and takes the grade 4 test four times, and get the scores 1300, 1400, 1650, and 1950 on the four occasions.
2. We could certainly say that Student A's scores show growth.
3. Note that with these scores of 1300, 1400, 1650, and 1950, Student would be categorized respectively at achievement Level 1, Level 2, Level 3, and Level 4.
4. Most of us would say that, since Student A took the grade 4 test four times, the increasing scores means that Student A learned more of the grade 4 content standards ("grade level expectations").

5. Now consider if Student A the next year were promoted to grade 5, and took the grade 5 test, and for some reason took it four times, and again got the scores 1300, 1400, 1650, and 1950. These scores happen to map to Levels 1, 2, 3, and 4, respectively—but at Grade 5.
6. In other words, the same student could take two different tests—Grade 4 and Grade 5—and get exactly the same (scale) scores, and get exactly the same Achievement Levels—but once in grade 4 and once in grade 5.
7. Consider a third scenario where the student again takes four tests, but this time takes the Grade 4, Grade 5, Grade 6, and Grade 7 tests. Again the student scores 1300, 1400, 1650, and 1950. In this set of tests and scores, the student would again be placed in achievement Level 1, Level 2, Level 3, and Level 4 for each of the different grade tests.

These three scenarios are summarized below.

Three Students, Same Scale Scores and Achievement Levels, Different Combinations of Tests					
	1300 (Level 1)	1400 (Level 2)	1650 (Level 3)	1950 (Level 4)	Interpretation of Growth
Student A	Grade 4	Grade 4	Grade 4	Grade 4	
Student B	Grade 5	Grade 5	Grade 5	Grade 5	
Student C	Grade 4	Grade 5	Grade 6	Grade 7	

Grade	Level 1	Level 2	Level 3	Level 4	Level 5
<b>3</b>	86 - 1045	1046 - 1197	1198 - 1488	1489 - 1865	1866 - 2514
<b>4</b>	295 - 1314	1315 - 1455	1456 - 1689	1690 - 1984	1965 - 2638
<b>5</b>	474 - 1341	1342 - 1509	1510 - 1761	1762 - 2058	2059 - 2713
<b>6</b>	539 - 1449	1450 - 1621	1622 - 1859	1860 - 2125	2126 - 2758
<b>7</b>	671 - 1541	1542 - 1714	1715 - 1944	1945 - 2180	2181 - 2767

What does each sequence of scores on the given tests mean in terms of what the student knows and can do?

Given that the scale scores are the same, under what conditions would we say that the growth of Student A is the same as the growth of Student B and Student C?

What tools and criteria do we have that will help us evaluate whether these conditions have been met? What tools and criteria do we still need to develop to help design assessments that will do what we want them to in terms of supporting inferences about student growth?

What are the implications for how tests—and scales and achievement levels—are designed and administered for out-of-level testing, growth-toward-a-proficiency-standard, and computer-adaptive testing?