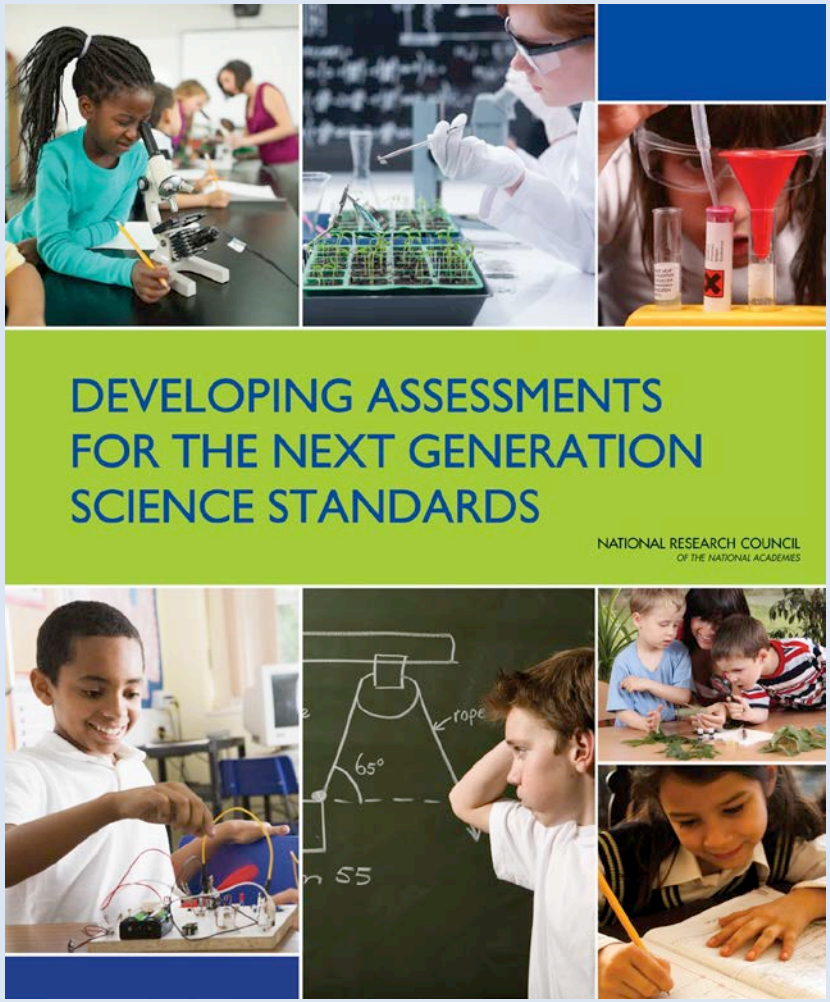
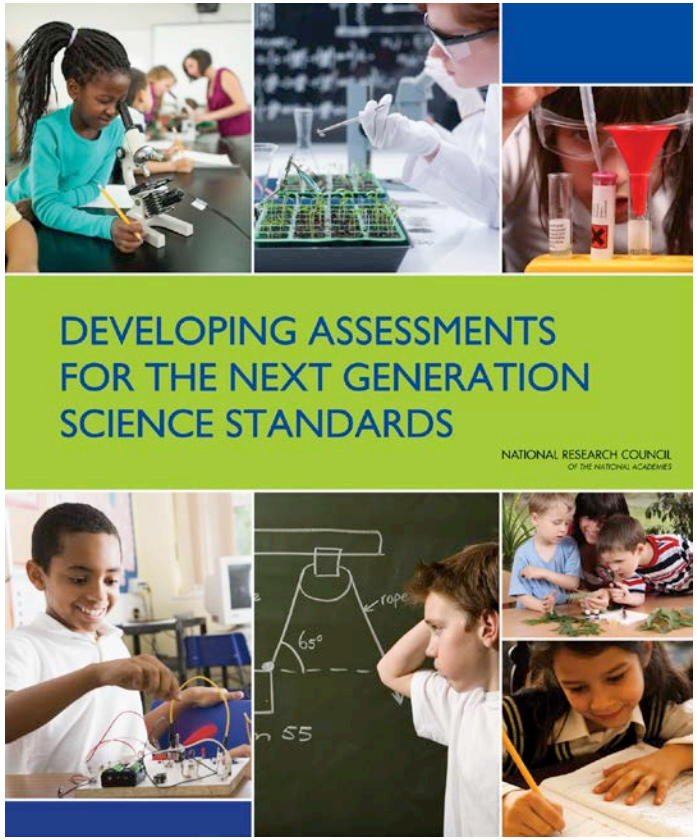
Measuring What Matters: Opportunities & Challenges in Assessing Science Proficiency



Jim Pellegrino

UIC LEARNING SCIENCES RESEARCH INSTITUTE

Large-Scale Monitoring Assessment: Design Challenges

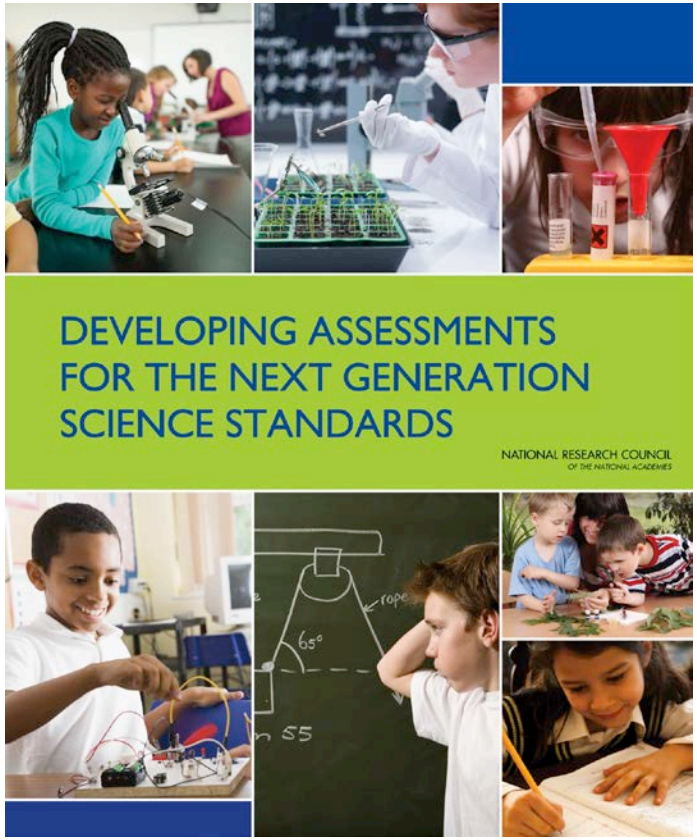


Challenge #1: Design of the Monitoring Component – Intended uses of the Information

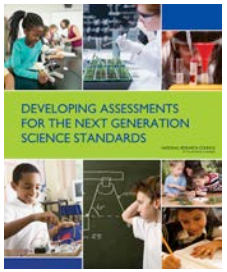
Complex Space of Monitoring Functions

TABLE 5-1 Questions Answered by Monitoring Assessments

Types of inferences	Levels of the Education System: Individual Students	Schools or District	Policy Monitoring	Program Evaluation
Criterion-referenced	Have individual students demonstrated adequate performance in science?	Have schools demonstrated adequate performance in science this year?	How many students in state X have demonstrated proficiency in science?	Has program X increased the proportion of students who are proficient?
Longitudinal and comparative across time	Have individual students demonstrated growth across years in science?	Has the mean performance for the district grown across years? How does this year's performance compare to last year's?	How does this year's performance compare to last year's?	Have students in program X increased in proficiency across several years?
Comparative across groups	How does this student compare to others in the school/state?	How does school/ district X compare to school/district Y?	How many students in different states have demonstrated proficiency in science?	Is program X more effective in certain subgroups?



Challenge #2: Design of the Monitoring Component – Possible Sources of Evidence



State Assessments for Monitoring

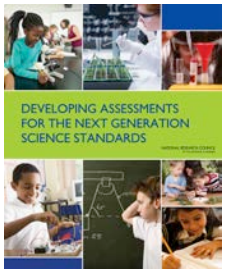
Combine two types of external assessment strategies, in conjunction with OTL indicators:

On-Demand Assessments

- Developed by the state
- Administered at a time mandated by the state

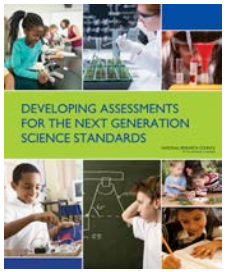
Classroom-Embedded Assessments

- Developed by the state or district,
- Administered at a time determined by the district/school that fits the instructional sequence in the classroom



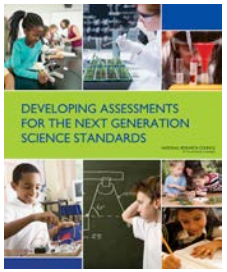
Options for On-Demand Assessments

- Mixed item formats, including extended constructed response
  - Such as the AP Biology
- Mixed item formats with performance tasks
  - might involve both group and independent activities (NECAP example)
  - might involve some hands-on tasks, such as having students perform tasks at stations (NY example)
- Use matrix sampling, depending on the intended use and the need to report scores for individuals versus for groups.



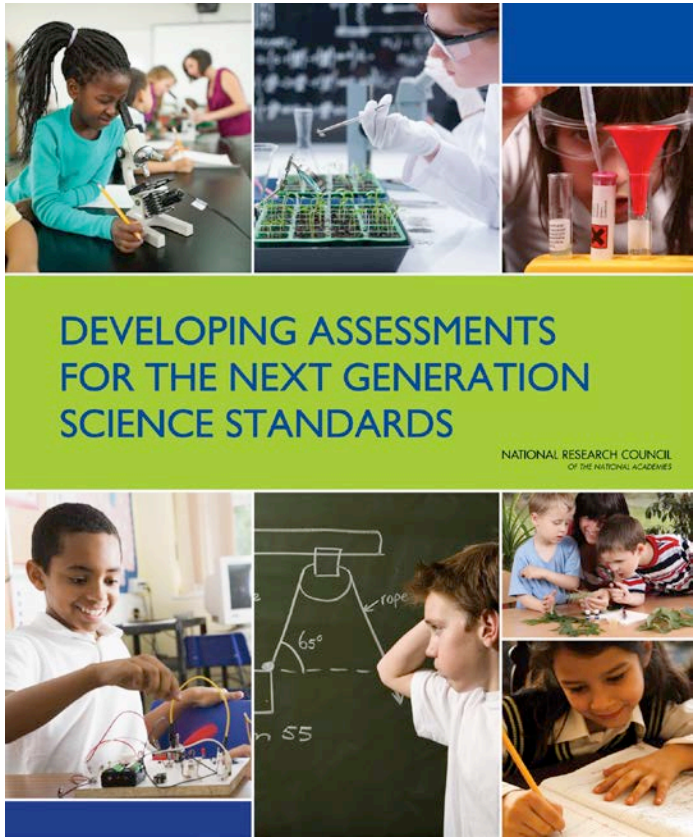
Possible Options for Classroom-Embedded Assessment Component

- **Replacement units** (curriculum materials + assessments) developed outside of the classroom (by state or district)
- **Item banks of tasks**, developed outside of the classroom
- **Portfolio collections of work samples**, with tasks specified by state or district

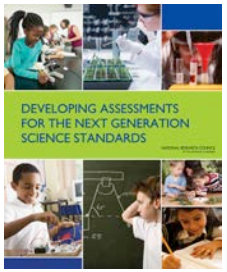


Classroom-Embedded Monitoring Assessments (cont.)

- Teachers administer them at a time that fits with their local curriculum sequence, possibly set by the school or district.
- Teachers receive training in how to administer the assessment(s)
- Scoring done by trained teachers or sent to the district/state for scoring
- Moderation and quality control procedures to enhance score comparability for desired inferences/comparisons needed for a monitoring purpose.

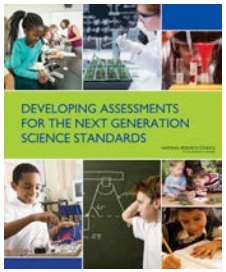


Challenge #3: Design of the Monitoring Component – Choice of Grade Levels & Representation of the Standards



Targeted Grade Levels & Targeted Standards

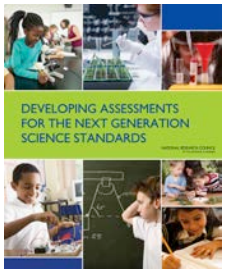
- ESSA requires assessment of science once at each of three grade bands
  - 3-5; 6-8; HS
- NGSS are specified in multiple ways
  - For K-5 they are written as grade level specific
  - For 6-8 & HS they are written in terms of the grade band
    - For instruction in 6-8 or HS the grade band "content" can be covered using an "integrated" model or a "discipline-specific" model
    - States can express a preferred model



Distributions of the Performance Expectations

Grade 5

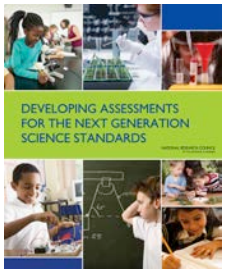
		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		1	2		1				4	
	2							1		1	
	3		1							1	
	4									0	6
LS	1							1		1	
	2		1							1	
	3									0	
	4									0	2
ESS	1				1			1		2	
	2		1			1				2	
	3								1	1	5
ETS	1	1		1			1			3	
	2									0	3
		1	4	3	1	2	1	3	1	16	



Distributions of the Performance Expectations

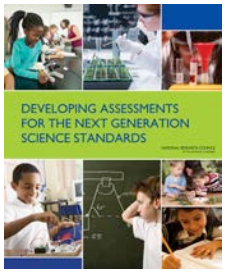
Grades 6 - 8

		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		3		1		1		1	6	
	2	2		4			2	2		10	
	3		2		2					4	
	4		2			2			2	6	26
LS	1		1				1			2	
	2		1		1		1	2		5	
	3		2							2	
	4				4	2	4		2	12	21
ESS	1		4		2		2			8	
	2		1		1		1			3	
	3				1		1	2		4	15
ETS	1	1	1		1			1		4	
	2									0	4
		3	17	4	13	4	13	7	5	66	



Targeted Grade Levels & Targeted Standards

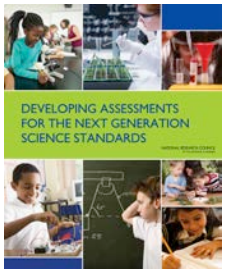
- The selected grade brings with it implications for choosing what is represented on the assessment at that grade
  - The grade level standards and/or the grade band standards?
- The selected grade brings with it implications for choosing who is being held accountable for what
  - The student (grade level or grade band)
  - The teachers at the school
    - at the target grade and/or for the grade band



Distributions of the Performance Expectations

Grade 5

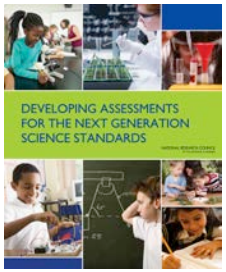
		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		1	2		1				4	
	2							1		1	
	3		1							1	
	4									0	6
LS	1							1		1	
	2		1							1	
	3									0	
	4									0	2
ESS	1				1			1		2	
	2		1			1				2	
	3								1	1	5
ETS	1	1		1			1			3	
	2									0	3
		1	4	3	1	2	1	3	1	16	



Distributions of the Performance Expectations

Grades K - 5

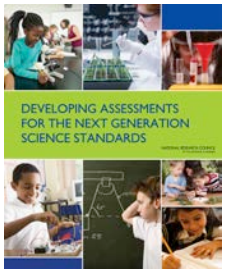
		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		1	3	1	1	1	1		8	
	2	2		3	1			1		7	
	3	1	1	2			3			7	
	4		2	2			3			7	29
LS	1		1		1		1	3	1	7	
	2		2	1				1		4	
	3				1		2			3	
	4			1	1		1	2		5	19
ESS	1			1	2		2	1		6	
	2		2	1	3	1	1	1	2	11	
	3	1	1				1	1	3	7	24
ETS	1	2	1	1	1		1			6	
	2									0	6
		6	11	15	11	2	16	11	6	78	



Distributions of the Performance Expectations

Grade 8

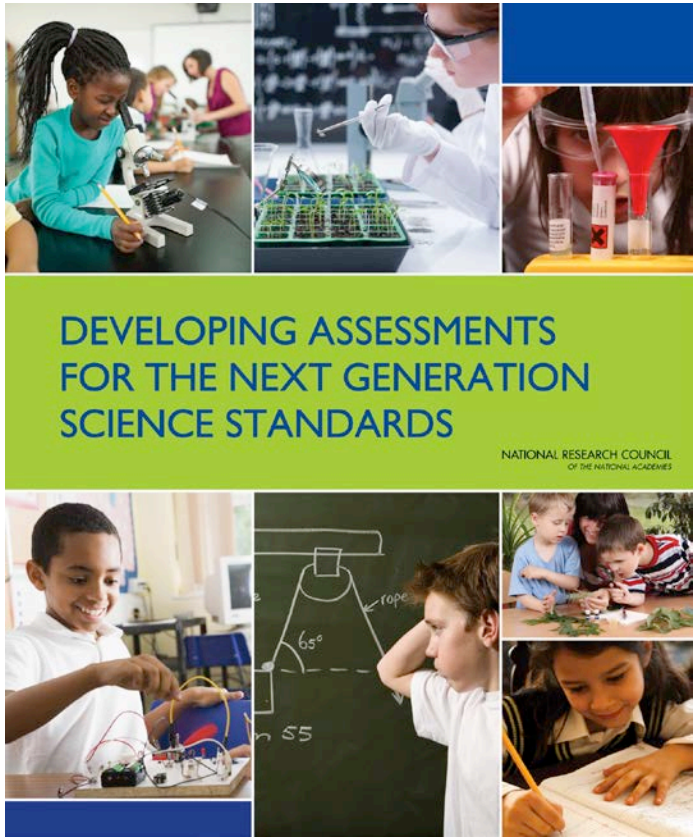
		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1									0	
	2	1		2			1	1		5	
	3		1		1					2	
	4		1			1			1	3	10
LS	1									0	
	2									0	
	3		1							1	
	4				2	1	2		1	6	7
ESS	1		2		1		1			4	
	2									0	
	3							1		1	5
ETS	1	1	1		1			1		4	
	2									0	4
		2	6	2	5	2	4	3	2	26	



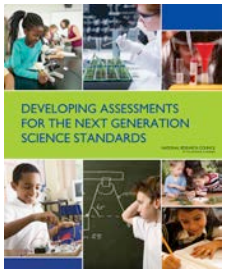
Distributions of the Performance Expectations

Grades 6 - 8

		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		3		1		1		1	6	
	2	2		4			2	2		10	
	3		2		2					4	
	4		2			2			2	6	26
LS	1		1				1			2	
	2		1		1		1	2		5	
	3		2							2	
	4				4	2	4		2	12	21
ESS	1		4		2		2			8	
	2		1		1		1			3	
	3				1		1	2		4	15
ETS	1	1	1		1			1		4	
	2									0	4
		3	17	4	13	4	13	7	5	66	

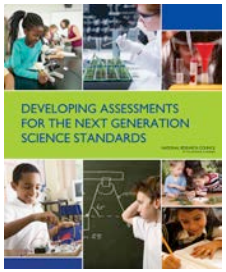


Challenge #4: Design of the Monitoring Component – Alignment with what – The NGSS' stated PE's or the "Matrix"



Alignment of Tasks with What: Stated PEs or the Matrix

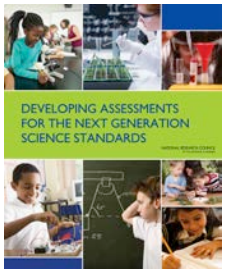
- As noted, the selected grade for administering the assessment brings with it implications for choosing what is represented on the assessment at that grade
  - The grade level standards and/or the grade band standards
- Either way, the matrix of PEs is "sparse"
  - Must tasks be "aligned" to only those cells specified
  - Can tasks be aligned to other possible intersections
    - Can the assessment contain a "mixture"



Distributions of the Performance Expectations

Grade 5

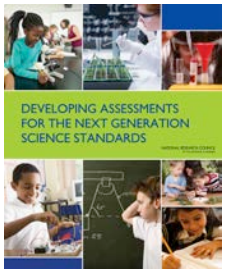
		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		1	2		1				4	
	2							1		1	
	3		1							1	
	4									0	6
LS	1							1		1	
	2		1							1	
	3									0	
	4									0	2
ESS	1				1			1		2	
	2		1			1				2	
	3								1	1	5
ETS	1	1		1			1			3	
	2									0	3
		1	4	3	1	2	1	3	1	16	



Distributions of the Performance Expectations

Grades 6 - 8

		SEP1	SEP2	SEP3	SEP4	SEP5	SEP6	SEP7	SEP8		
PS	1		3		1		1		1	6	
	2	2		4			2	2		10	
	3		2		2					4	
	4		2			2			2	6	26
LS	1		1				1			2	
	2		1		1		1	2		5	
	3		2							2	
	4				4	2	4		2	12	21
ESS	1		4		2		2			8	
	2		1		1		1			3	
	3				1		1	2		4	15
ETS	1	1	1		1			1		4	
	2									0	4
		3	17	4	13	4	13	7	5	66	



Figuring Out What to Assess: Claims, Evidence, Tasks & Alignment

- You have to start with the domain and how you are choosing to define it
  - Grade Level or Grade Band; PEs, Bundles of PEs, or ????
- Whatever your Domain definition, you have to figure out how to represent it and sample from within it
- PEs and/or Bundles of PEs encompass complex sets of knowledge and skills – the three dimensions
- You can't assess it all – but what you do assess should be "significant" aspects of the critical 3-D knowledge
- You have to have a way to specify what are "significant" parts of the knowledge space – whether you are building stand alone tasks or task clusters
- You need a principled design process to figure this out