Wyoming Assessment Task Force: Meeting #3 via WebEx

Joseph Martineau & Scott Marion, Center for Assessment

via WebEx

Center for Assessment

July 13, 2015

WebEx Protocol

- Last time we did not have the ability to involuntarily mute or log people out. That has been resolved. We will make sure that this WebEx is a better experience than the last one.

- If you called the WebEx rather than having the WebEx call you, please log out and log back in using the option to have the WebEx call you.

Attendance

Present	Name	Present	Name
	Dan Coe	Y	Kevin Mitchell
N	Stephanie Czarobski	Y	Anne Ochs
Y	Sharla Dowding	Y	Mary Charles Pryor
Y	Christopher Dresang	N/A	Jon Lever
Y	Kim Ferguson		Kevin Roberts
Y	Molly Foster		Sue Stevens
Y	Crystal Graf	Y	Byron Stutzman
Y	Cindy Gulisano	N	Sonya Tysdal
	Joanne Flanagan	Y	Kathy Vetter
Y	Shannon Harris	Y	Rebecca Weston
Y	Cassie Hetzel	Y	Nicole Novotny Wonka
Y	Ellen Kappus		Marty Wood
Y	Audrey Kleinsasser		
Y	Wanda Maloney		

Intuitive Test Theory

- P-prims (Braun and Mislevy, Intuitive Test Theory)
  - Did any of the p-prims of intuitive test theory resonate with you?
  - In your work with assessment, have you run up against intuitive p-prims?

- Another couple of intuitive p-prims most of us have heard
  - "Standardized assessments are invalid."
  - "Standardized assessments only tell me what I already know about my students."
  - Which one is it?

Intuitive Test Theory

- The utility of intuitive p-prims for traditional classroom interim and summative assessment situations can lead to
  - Oversimplification of standardized assessment
  - Belief that teachers are inherently good at applying sound assessment principles in their interim and summative assessment practices
  - Misunderstanding the definition of and complexities of formative assessment (for another day)
- A trap in debunking intuitive p-prims
  - *Multiple choice (MC) measures only recall* is an (inaccurate) intuitive p-prim
  - The trap: Just because MC can be used to measure knowledge and skill beyond simple recall does not mean that we don't need to go beyond multiple choice to better measure complex knowledge and skills.

Intuitive Test Theory

- Intuitive p-prims sometimes infect the work of assessment professionals.
  - Standardized interim assessments can inform instruction in the way teachers want it to inform instruction.
  - Item banks serve as daily instructional tools.
- Intuitive p-prims are regularly present in the marketing of assessment products.
  - Interim assessment products such as those developed by teachers, districts, states, consortia (PARCC, Smarter Balanced), or vendors (NWEA, Renaissance Learning, ACT) are sufficient for data-driven instruction.
  - Teachers can learn to implement high-quality assessment practices in a two-hour presentation (Pearson, Marzano), or even a week-long workshop

Principled Assessment Design

- What's that?
- Too often assessments are designed by simply trying to match test questions to individual standards or other criteria
  - This leaves us wanting in how to meaningfully interpret the results
    - How are students developing competence in the domain?
    - Do statistical techniques provide information that match the way that students learn?
  - *A corollary:* This causes us problems with justifying the claims we want to make based on the results of assessment
    - Isolating a content standard in a test question ignores the connected knowledge and skills we desire to represent

Principled Assessment Design

- Advances in measurement theory over the past 25-30 years has provided insights about how to create more useful test designs
  - Bob Mislevy and colleagues
  - *Knowing What Students Know* and other National Research Council publications
- Advances in measurement practice are slowly changing to incorporate these ideas
- Principled assessment design is difficult and costly work. Precursors are much easier and less costly.

Evidence Centered Design

- It is a somewhat complex theoretical approach created by Bob Mislevy and colleagues (1994, 1996), but its core is actually quite easy to understand

- This approach to task/assessment design is an attempt to explicitly connect advances in our understandings of learning with our deepening understanding of educational assessment

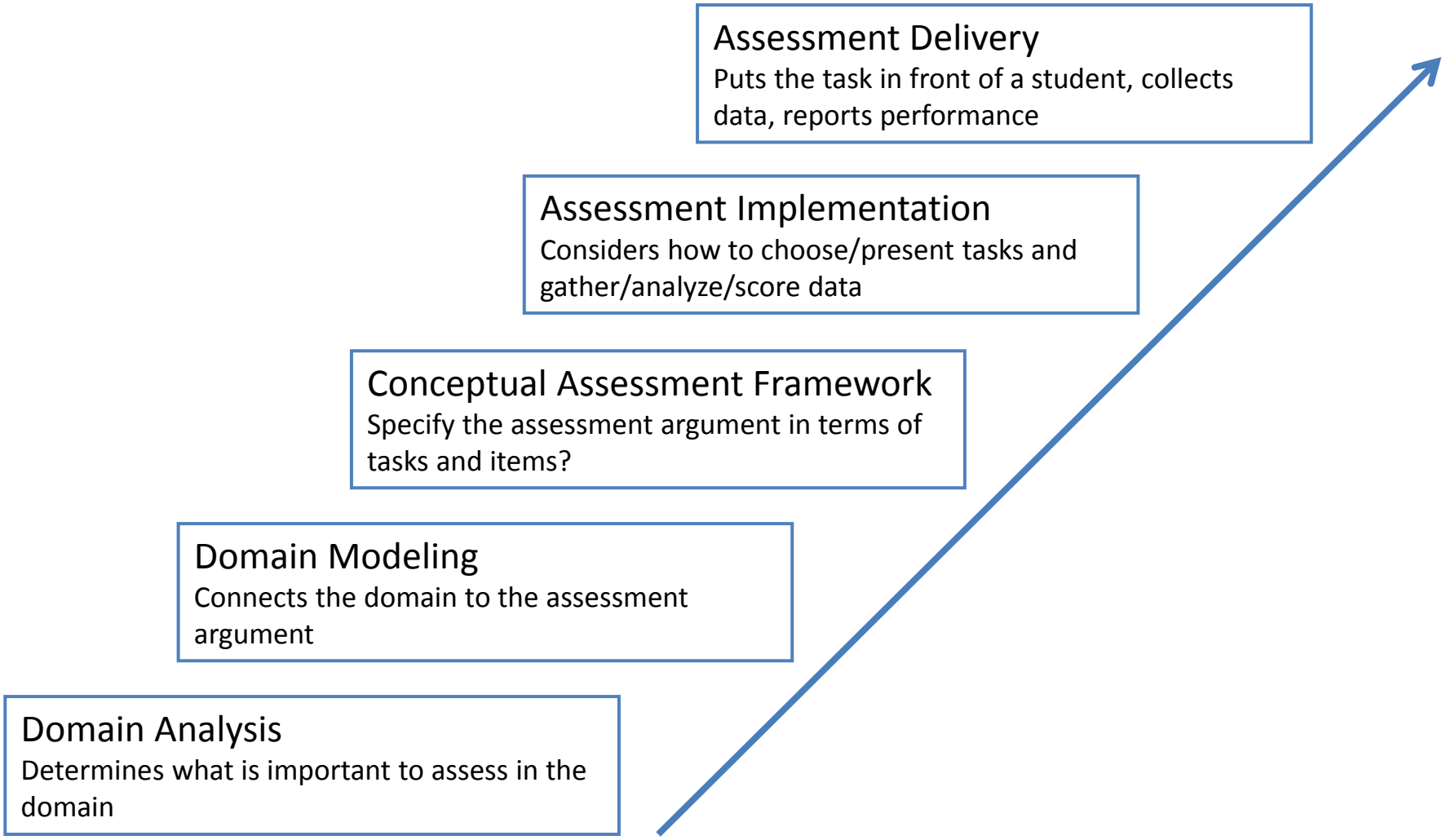
Principles of ECD

- Must engage in purposeful thinking and validation up front!
- Everything flows from and is defined by the purpose of the assessment and intended inferences
  - Remember, validity is about the evidence that you have to support inferences from test scores!
- Claims (what the designers are *claiming* to measure) are explicitly defined
- Design tools and items/tasks resulting from the assessment design and development process coherently relate back to the purpose of the assessment
- There is clear documentation and explanation of the evidence/rationale for each decision made within the assessment design process (e.g., item format, measurement model applied)
- The design process is iterative and provides for the revision of tools as deemed appropriate/necessary

Overview of ECD

- **Student model**—exactly what do you want students to know and how (well) do we want them to know it? This requires a very careful examination of the "construct" to unpack the thing we want students to know and how well we want them to know it. It is NOT just the standards!!!
- **Evidence model**—what will you accept as evidence that the student has the desired knowledge? This is really like a thought experiment where one needs to describe what sort of evidence would convince you that the student demonstrated the knowledge and skills described in the student model.
- **Task model**—what tasks will students perform to demonstrate/communicate their knowledge? Once we have figured out the evidence that would convince us that the student has learned what was intended, can we design a task or tasks that would elicit that evidence.

5 Layers of ECD Framework



Knowing What Students Know (2001)

- One of the most important assessments reports to ever come out of the National Research Council
- It builds off of Mislevy's work but presents a very clear message about the importance of ensuring that assessments are coherent with the underlying model of learning

KWSK: Assessment as a Process of Reasoning from Evidence

- Cognition

  model of how students represent knowledge & develop competence in the domain

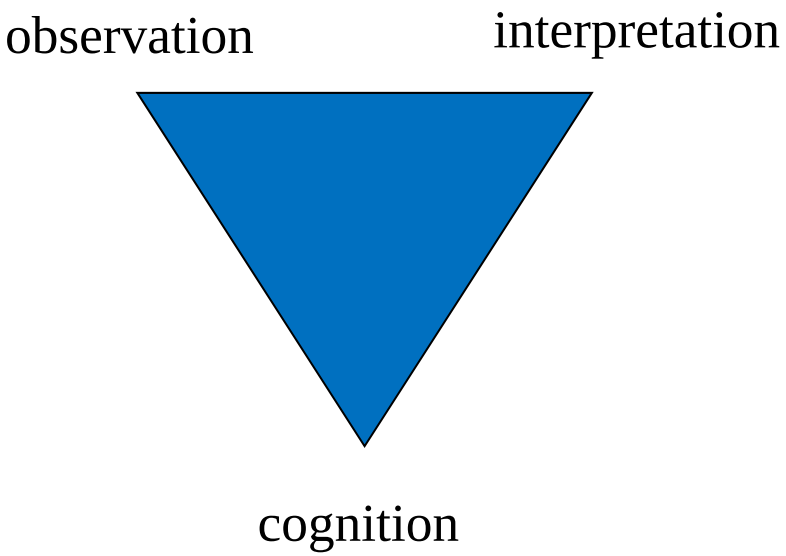
- Observation

  tasks or situations that allow one to observe students' performance

- Interpretation

  method for making sense of the data

The Assessment Triangle



Principled Assessment Design: *KWSK*

- NRC's (KWSK) Assessment Triangle, continued…
  - *Cognition:* We need to be thoughtful about what our learning model is so that we can define the domain of knowledge and skills (our content standards) in a way that provides a structure for connecting the standards to the complexity, progression, and interconnectedness of the learning model.
  - *Observation:* Once we have a clear model and set of claims about what and how students are expected to achieve competence in the domain, it is then possible to develop tasks and problems that also match the learning model, and connect the assessment to the high-quality instruction.
  - *Interpretation*: It is then possible to apply appropriate techniques in order to interpret test scores in light of the learning model and have confidence that the test scores are valid measures of learning under modern instructional practices.
  - If any of these three legs of the stool is missing, the validity of test scores for their intended purposes is questionable

Principled Assessment Design

- The use of principled assessment design is exceedingly rare
  - Advanced Placement exams that have been revised in recent years
  - PARCC
  - Smarter Balanced
  - NCSC

- We provide two quick examples from AP and PARCC (SBAC followed a similar process)

College Board – AP Exams

- Applied Measurement in Education, Volume 23 (2010)
- Followed an ECD-based approach to specify targets of instruction for AP courses and targets of measurement for AP assessments
  - Subject matter experts performed a *Domain Analysis*
    - Identify and prioritize important content in the domain; define specific skills necessary to interact with that content
  - Domain Modeling: Generation of claims and evidence statements and ALDs
    - Focus on alignment between curriculum, instruction, assessment and reporting

AP Domain Analysis

TABLE 1
An Example Content Outline in Chemistry for one Big Idea

Big Idea: Changes in matter involve the rearrangement and/or reorganization of atoms and/or the transfer of electrons.

Enduring Understanding: Chemical changes are represented by a balanced chemical reaction that identifies the ratios with which reactants react and products form.

Supporting Understandings:

A.1 A chemical change may be represented by a molecular, ionic, or net ionic equation.

A.2 Quantitative information can be derived from stoichiometric calculations which utilize the mole ratios from the balanced equations. ***(Possible examples:*** the role of stoichiometry in real world applications is important to note so that it does not seem to be simply an exercise done only by chemists; and the concept of fuel-air ratios in combustion engines, for example, is able to provide context for this form of calculation.)

A.3 Solid solutions, particularly of semiconductors, provide important, non-stoichiometric compounds. These materials have useful applications in electronic technology and provide an important extension of the concept of stoichiometry beyond the whole number mole-ratio concept.

AP Domain Analysis

TABLE 2
Sample Skills and Skill Definitions from Science

1. **Evaluate scientific questions**
   - 1A. Justification that question is in scope of investigation and domain
   - 1B. Evaluation and criteria for the evaluation appropriate to the question
   - 1C. Specification of causal mechanism(s) that is related to the question
   - 1D. Validity of the claim that the focus of the question is related to its purpose
2. **Apply mathematical routines to quantities that describe natural phenomena**
   - 2A. Appropriateness of application of mathematical routine in new context
   - 2B. Appropriateness of selected mathematical routine
   - 2C. Correctness of mapping of variables and relationships to natural phenomena
   - 2D. Correctness of application of mathematical routine
   - 2E. Correctness of results of mathematical routine
   - 2F. Reasonableness of solution given the context
   - 2G. Description of the dynamic relationships in the natural phenomena
   - 2H. Prediction of the dynamic relationships in the natural phenomena
   - 2I. Precision of values consistent with context
3. **Connect concepts in and across domain(s) to generalize or extrapolate in and/or across enduring understandings and/or big ideas.**
   - 3A. Articulation of content-specific relationships between concepts or phenomena
   - 3B. Prediction of how a change in one phenomenon might effect another
   - 3C. Comparison of salient features of phenomena that are related
   - 3D. Appropriateness of connection across concepts
   - 3E. Appropriateness of connection of a concept among contexts

College Board - AP Exams (cont.)

- Claims were combinations of individual content and skill features resulting from domain analysis.
  - Much discussion around appropriate grain size
  - Target proficiency level - end of course
- Evidence statements
  - What students need to do to show acquisition of claim
  - "The work is characterized by…"
- Use of claims and evidence to generate ALDs
- Use ALDs, claims and evidence to support the development of task templates design

PARCC use of ECD

- Common Core State Standard
- Content Model Frameworks
- Claims
- Evidence Statements – define evidence that will be collected in support of a given sub-claim
- Test Blueprints - assessment targets/conditions for assessment
- Performance Level Descriptors
- Design Patterns
- Item and Task Generation Model

Claims Structure: Mathematics

Master Claim: On-Track for college and career readiness. The degree to which a student is college and career ready (or "on-track" to being ready) in mathematics. The student solves grade-level /course-level problems in mathematics as set forth in the Standards for Mathematical Content with connections to the Standards for Mathematical Practice.

Total Exam Score Points: 92 (Grades 3-8), 107 (HS)

0

Sub-Claim A: Major Content[1] with Connections to Practices

The student solves problems involving the Major Content[1] for her grade/course with connections to the Standards for Mathematical Practice.

~40 pts (3-8), ~50 pts (HS)

Sub-Claim B: Additional & Supporting Content[2] with Connections to Practices

The student solves problems involving the Additional and Supporting Content[2] for her grade/course with connections to the Standards for Mathematical Practice.

~18 pts (3-8), ~25 pts (HS)

Sub-Claim C: Highlighted Practices MP.3,6 with Connections to Content[3] (expressing mathematical reasoning)

The student expresses grade/course-level appropriate mathematical reasoning by constructing viable arguments, critiquing the reasoning of others, and/or attending to precision when making mathematical statements.

14 pts (3-8), 14 pts (HS)

Sub-Claim D: Highlighted Practice MP.4 with Connections to Content (modeling/application)

The student solves real-world problems with a degree of difficulty appropriate to the grade/course by applying knowledge and skills articulated in the standards for the current grade/course (or for more complex problems, knowledge and skills articulated in the standards for previous grades/courses), *engaging particularly in the Modeling practice*, and where helpful making sense of problems and persevering to solve them (MP. 1),reasoning abstractly and quantitatively (MP. 2), using appropriate tools strategically (MP.5), looking for and making use of structure (MP.7), and/or looking for and expressing regularity in repeated reasoning (MP.8).

12 pts (3-8), 18 pts (HS)

Sub-Claim E: Fluency in applicable grades (3-6)

The student demonstrates fluency as set forth in the Standards for Mathematical Content in her grade.

7-10 pts (3-8)

[1] For the purposes of the PARCC Mathematics assessments, the Major Content in a grade/course is determined by that grade level's Major Clusters as identified in the *PARCC Model Content Frameworks* for Mathematics (with designations for high school courses to come in the final Frameworks). Note that tasks on PARCC assessments providing evidence for this claim will sometimes require the student to apply the knowledge, skills, and understandings from across several Major Clusters.

[2] The Additional and Supporting Content in a grade/course is determined by that grade level's Additional and Supporting Clusters as identified in the *PARCC Model Content Frameworks* for Mathematics.

[3] For 3 – 8, Sub-Claim C includes only Major Content. For High School, Sub-Claim C includes Major, Additional and Supporting Content.

Claims Driving Design: ELA/Literacy

Students are on-track or ready for college and careers

Students read and comprehend a range of sufficiently complex texts independently

Students write effectively when using and/or analyzing sources.

Students build and present knowledge through research and the integration, comparison, and synthesis of ideas.

Reading Literature
RL.X.1-10

Reading Informational Text
RI.X.1-10 and Reading Literacy Standards

Vocabulary Interpretation and Use
RL/RI.X.4 and L.X. 4-6

Written Expression
W.X.1-10 and Disciplinary Writing Standards

Conventions and Knowledge of Language
L.X.1.-3

Questions?

- So why do we care about all this fancy measurement stuff anyhow?

- Aren't we just figuring out how to pick test questions to measure the standards?

Drivers of Design Decisions

- Learning Model
  - *Cognition* leg of the assessment triangle (how do we believe that knowledge and skills are acquired)
  - *Intuitive p-prim*
    - Theory of learning is not critical to assessment design as long as we clearly define our content standards and pick items to measure them
  - *Complex p-prim*
    - Theory of learning affects the way in which we understand how students progress through a complex domain
      - In an *atheoretical* approach, we are simply left with the "1000 mini-lessons problem"
      - A theoretical (or principled) approach where the assessment is designed according to a specific learning model, users are able to see where learners are in a complex map of the domain
      - Further, a clear conception of learning is critical for connecting aspects of a comprehensive assessment system

Drivers of Design Decisions

- Content Standards
  - *Intuitive p-prim*
    - An on-demand test is too short to get at all of the content standards: randomly sampling from the standards to represent overall knowledge of the content standards as a whole is a generally viable approach
  - *Complex p-prim*
    - The structure of content standards determines how to adequately represent the content standards in a relatively short test, for example…
      - With *atheoretical* content standards, random sampling of the content standards is adequate to *claim* that performance on the test represents achievement on the overall collection of discrete knowledge and skills. This has two considerable problems
        - » Random sampling leaves some standards un-measured, and a single random draw can randomly leave out some of the most important knowledge and skills
        - » This approach does not allow for important connections between content standards because content standards are randomly sampled
      - With *constructivist-style* content standards, purposive care is required in specifying combinations of skills and understanding that support a claim that students can apply and synthesize understanding of foundational principles to construct novel solutions

Drivers of Design Decisions

- Purposes and Uses
  - Inform instruction
    - *Intuitive p-prim*
      - Interim assessment data facilitates data-driven instruction
    - *Complex p-prim*
      - Interim assessment occurs too infrequently to support instructional decisions
      - Assessment information to support instruction requires careful consideration of:
        - » Timing (related to curriculum and instruction)
        - » Types of items/tasks (designed to provide summary information or insight into student learning)
        - » Nature of content (focus, depth, and breadth)
        - » Form of the results and feedback (summaries, descriptions)
        - » Level of support

Drivers of Design Decisions

- Purposes and Uses
  - Measure student growth
    - *Intuitive p-prim*
      - Growth is the post-score minus the pre-score as long as both tests are from the same program (e.g., NWEA, ACT, Smarter Balanced, PARCC, PAWS, etc.)
    - *Complex p-prim*
      - The pre- and post-test have to be on the same scale with very similar blueprints for simple subtraction to work.
      - Even if the conditions are met, it is unclear what the difference score means in terms of what new content a student learned.
      - If we want growth scores to convey meaning about what a student has learned, traditional assessment design is insufficient.

Questions?

- This is just a sampling of the uses and purposes.

- Any questions, comments, or discussion about the sampling we presented?
- What other uses and purposes would you like to ask about, make a comment about, or discuss?

Critical Design Considerations

- The learning model, content standards, and purposes and uses provide a framework for considering critical design decisions.
- We said “Drivers of Design Considerations” but that’s not entirely true.
- Design considerations will be strongly influenced by this framework, but will not necessarily drive the decisions; there are other considerations such as legislative requirements, cost, and time constraints.
- Each design decision has implications, often a whole set of implications.
- The implications often are about how well the test matches the learning model, content standards, and purposes and uses.
- There are important practical implications as well such as burden on districts, schools, teachers, and students.
- .

Critical Design Considerations (CDCs)

- We need to recognize that we cannot hope to make every design decision perfectly consistent with the "Drivers" of Design Considerations
- Based on this foundation, we start with the most important design consideration:

  What claims do we desire to make based on the results of the assessment?

- This question will strongly influence most other design considerations, and derives directly from the learning model, content standards, and intended purposes and uses.

CDC: Intended Claims

- Implications
  - The claims we want to make certain kinds of tasks more appropriate than others to support the desired claims, for example…
    - Multiple Choice
    - Long essay
    - Performance Tasks
  - More complex claims will be more expensive to support
    - Simpler items are less expensive
    - Handscoring is one of the most expensive part of state summative assessment programs
  - But…
    - Intuitive p-prim from *Intuitive Test Theory*: MC only measures recall
    - A corollary intuitive p-prim: PTs measure higher level skills

CDC: Intended Claims

- Implications, continued…
  - More complex claims about interconnected, deeper knowledge and skill will require more time for assessment
    - More complex task types take longer for students to complete and result in a comparatively small number of data points
    - Likely to support "less reliable, more valid" claims
  - Less complex claims about discrete, surface skills require less time for assessment
    - Less complex task types take less time to complete and result in a lot of data points quickly
    - Likely to support "more reliable, less valid" claims
  - More on this tension later.

Questions?

- Identifying intended claims is the most critical of the design considerations.

- Any questions, comments, or discussion about this design consideration?

Critical Design Considerations

- One critical design consideration that flows directly from the claims we want to make is how to convey information about those claims to our stakeholders
  - If the information is not easily understandable it is generally useless
  - If the information is not accurate, it can be damaging
- These two are often in conflict, as accuracy generally involves nuance
- *Addressed to some degree on page 15 of the Michigan report*

CDC: Conveying Information About Claims to Stakeholders

- Implications
  - We need to clearly understand who are the stakeholders and what are the desired vs. appropriate uses of data for those stakeholders
    - Do we include try to be inclusive of the full array of stakeholders?
      - Students and parents
      - Teachers, principals, district program & instructional staff, district superintendents, district school board members
      - Local legislative councils and mayors, state legislators and governor, state school board and superintendent
      - University faculty, institutional researchers, and admissions staff
      - Local business owners, state business leaders
      - News organizations and the general public
    - Do we prioritize our efforts? If so, how?

CDC: Conveying Information About Claims to Stakeholders

- Implications, continued…
  - To what degree do we sacrifice nuance and (and therefore accuracy) to convey information about claims to critical stakeholders?
  - To what degree do we invest in "invisible" professional development to avoid sacrificing nuance, knowing that even so, more nuanced information will be misused by some key stakeholders?
  - No matter the audience, reports and data files are likely to be over- and misinterpreted
    - How do we encourage more use of data for making important decisions while discouraging over- and misinterpretation
    - How do we convey the limitations of data and reports without stakeholders dismissing the program as "useless"

CDC: Conveying Information About Claims to Stakeholders

- Implications, continued…
  - **No matter the audience, if the answer to a question is not easily available, information is unlikely to be used to appropriately inform decisions**
    - Do we build expensive tools to easily guide various stakeholders through appropriate tabular and graphical information displays to answer their questions?
    - How do we guide stakeholders from the answer to one question to finding answers to the most common next logical questions (e.g., drilling down and summarizing)?
  - **The more claims we want to make and the more information about each claim that we want to convey, the longer the test will need to be to support those claims**
    - **Have to weigh testing time vs. reporting more than just an overall score and performance level**
    - **Have to weigh testing time vs. accurately representing the complex knowledge and skills and their interrelationships in the content standards**

Questions?

- Conveying information to stakeholders is another of the most critical design considerations

- Any questions, comments, or discussion about this design consideration?

Critical Design Considerations

- Another critical design consideration that flows directly from the claims we want to make is the alignment of the assessment to the content standards (and the learning model)
  - If the tasks on the assessment do not adequately represent the content standards, the claims will not be supportable
  - There are several aspects of alignment that are important in assessment design

- *Addressed to some degree on pages 7 and 19 of the Michigan report*

CDC: Alignment

- Implications
  - If we can't measure everything in every statement included in the content standards in an on-demand statewide summative assessment of reasonable length, we have to have a way to prioritize
    - Currently three broad methods:
      - Domain sampling approach
      - Federal peer review approach
      - ECD-based approach
    - We review each briefly below

CDC: Alignment

- Domain sampling approach
  - The oldest and most common approach on commercial assessments
  - Generally based on a domain of discrete statements of knowledge and skills
  - Allows for random sampling of the knowledge and skill statements to avoid writing tasks to test every statement
  - If the sampling is random, it is by definition statistically representative of the whole domain
  - Major weaknesses are:
    - A single random draw from the domain can easily result in non-representation of specific types of content
    - Does not deal with interconnections of content knowledge and skill statements
    - Does not deal well with achieving appropriate representation of cognitive complexity
    - Easy to justify dropping the most difficult to measure content knowledge and skills because it is a sample anyway

CDC: Alignment

- Federal peer review approach
  - Relatively recent extension of domain sampling, nearly universal on existing state summative tests
  - Generally based on a domain of loosely connected statements of knowledge and skills, with very rudimentary progressions across grades, and with associated designations of cognitive complexity required by each statement
  - Every statement is represented rather than sampling from the whole domain
  - Added a new requirement that the cognitive complexity of the standards and the test questions had to also be analyzed

CDC: Alignment

- Federal peer review approach
  - Representing every statement causes major problems with test length
  - As background, content standards are usually organized hierarchically, for example

Subject	Science	S
Strand	Biology	S.B
Domain	Genetics	S.B.G
Benchmark	DNA	S.B.G.A
Standard	Describe the structure of DNA	S.B.G.A.1
Standard	Describe the function of DNA	S.B.G.A.2

  - States took various approaches to the test length problem (from "Power Standards" to evaluating alignment at the Benchmark level instead of the Standard level, to covering all standards across all forms rather than on every form).
  - Remaining weaknesses:
    - Easy to marginalize the most difficult to measure (and likely higher-level) standards because we're dealing with prioritizing anyway
    - Does not well address connections between various content knowledge and skill statements

CDC: Alignment

- Principled Assessment Design Approach
  - The newest and least-common approach to achieving alignment
  - The importance of principled assessment design is that it allows for a highly-principled prioritization of connected content knowledge and skills without badly marginalizing those that are the most difficult to measure
  - Why is this important?
    - We have a growing body of research on the importance of using complex performance tasks instead of or in addition to traditional standardized testing as a way to prioritize certain high-leverage learning outcomes and connect with instruction
  - An Anecdote
    - Michigan Report (p. 7) concluded at the time it was issued that only two products had reasonably adequate documentation of alignment…

CDC: Alignment

- Implications
  - The fit of the approach used to document alignment to the content claims is critical to supporting those claims
  - The method of prioritizing to make test length manageable is critical to supporting the desired claims
    - Domain sampling is relatively inexpensive, peer-review alignment is more expensive, and principled assessment design is still more expensive. They are also least, more, and most time-consuming, respectively.
    - Avoiding the elimination of connectedness and hard-to-measure tasks is particularly costly in terms of funding and testing time on both the design and implementation sides
    - Except in my school (!!) “what gets tested gets taught” and “accountability drives the focus of instruction.” So, how do we balanced with cost and testing time?

Questions?

- Alignment is another of the most critical design considerations.

- Any questions, comments, or discussion about this design consideration?

Critical Design Considerations

- Another critical design consideration that flows directly from the claims we want to make is that the assessment is appropriate for students with disabilities, English learners, and for the various demographic groups that will take the test (bias for and against certain demographic groups)
  - Students' particular disabilities, lack of English fluency, and background knowledge particular to their specific demographic groups can get in the way of demonstrating their knowledge and skills if the assessment is not carefully designed
  - This is about the test itself getting in the way of some students demonstrating their knowledge and skills, or bias
  - If a student's score represents something less than their actual knowledge and skill because of disability, lack of English fluency, or lack of non-academic background knowledge available to other groups, states, districts, schools, and teachers serving certain demographics may be advantaged or disadvantaged with considerable consequences

- *Addressed to some degree on pages 6-9 and 34-35 of the Michigan report*

CDC: Bias

- Implications
  - To avoid as many problems up front, it is important to use Universal Design from the beginning, for example
    - Avoid irrelevant language load such as non-subject specific complex vocabulary, unnecessarily complex sentence structure
    - Avoid low-contrast graphics
    - Minimize the number of items that can't be produced in Braille
    - Avoid idioms and colloquialisms
  - Universal design can take us far, but there are limitations.
    - Accommodations (supports to reduce barriers to access) may be necessary for some students with disabilities and English language learners
    - While educators representing and/or with experience teaching certain demographic groups can be included from the beginning, potentially biasing content makes its way through

CDC: Bias

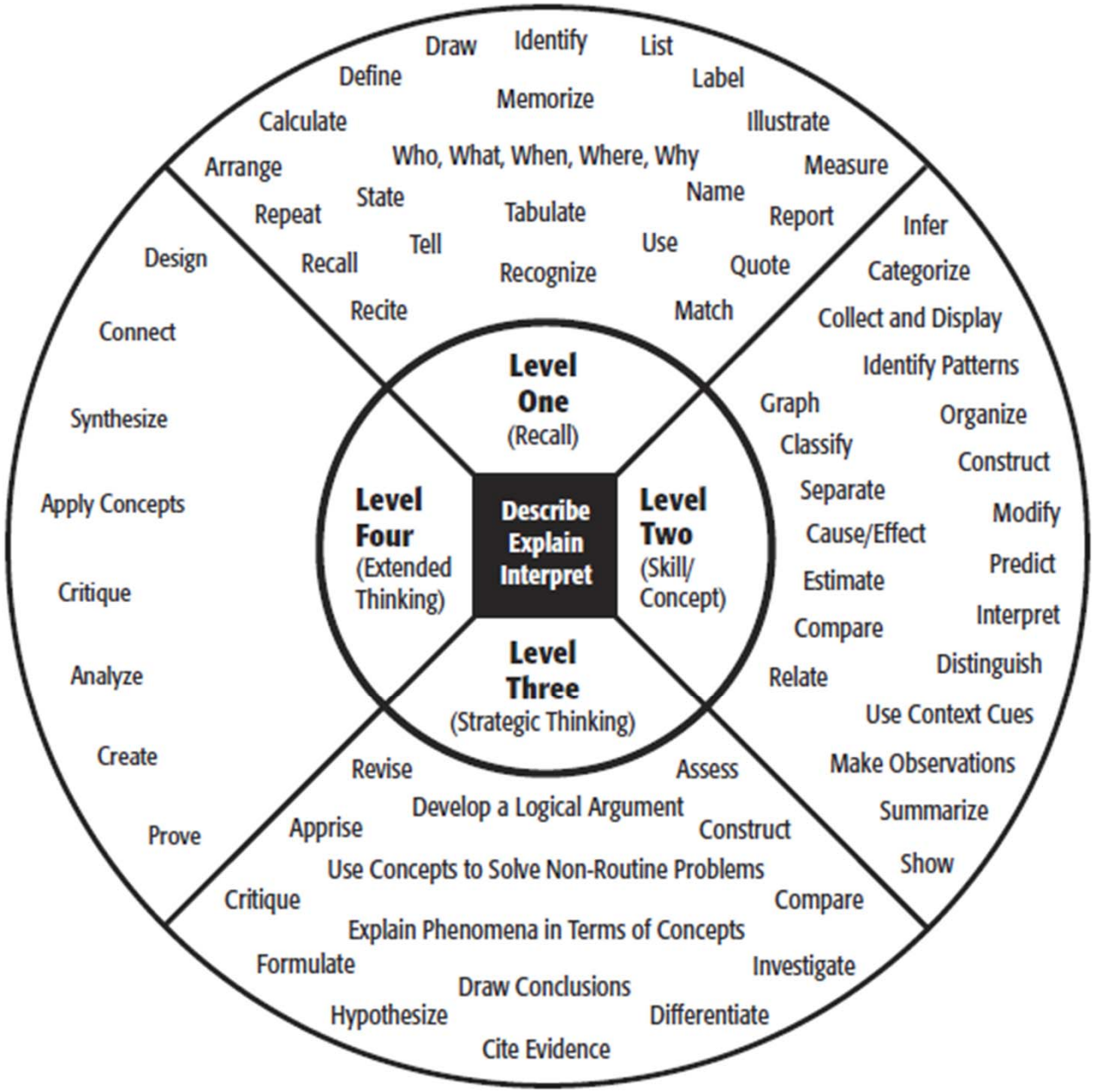
- Implications
  - We have to be careful to prevent bias from the start, and monitor for and eliminate bias throughout the entire process
    - Item writers live in their culture and with their background knowledge (sports, basements, attics, mowing the lawn…)
    - Culture and background knowledge are typically transparent to us (we don't see the air we breathe), and only become somewhat less transparent with training.
    - Even with detailed training, item writers can still unintentionally introduce cultural issues and background knowledge into the items they write
  - It is critical to monitor for and eliminate bias regardless of how diverse a population of students may be
    - Test content needs to be reviewed critically by specially trained independent people viewing the content through the lens of various cultural groups at multiple points throughout development
    - Even trained advocate-reviewers may not be able to spot all potential bias because they too live in their own (or adopted) culture and have a lifetime of accumulated background knowledge
    - Tasks need to be statistically analyzed to identify potential bias not caught through subjective review

Critical Design Considerations

- Another critical design consideration that flows directly from the claims we want to make is that the types of items and tasks appropriate to support those claims are suggested by them
  - With great care, simple task types can indirectly elicit evidence of complex thinking
  - More complex task types are more natively conducive to eliciting direct evidence
  - Just because an item type is complex does not necessarily mean it measures complex knowledge and skills

- *Addressed to some degree on pages 6-7 and 18-19 of the Michigan report*

Depth of Knowledge (DOK) Levels



From http://blog.curriculet.com/depth-of-knowledge-questions/

Critical Design Considerations

- Another critical design consideration that flows directly from alignment considerations is that the standards determine the degree to which items/tasks should be "naked" or highly-contextualized
  - Implication
    - Because of concerns with language load, the items/tasks need to be contextualized adequately, but no more than adequately
    - Highly connected content standards generally require adequate item/task contextualization, but Highly contextualized items/tasks are uncommon in commercial and state tests

Critical Design Considerations

- Closely related to item/task type and item/task contextualization are Performance Task design considerations

- *Addressed to some degree on pages 6-7, 12-13, and 16-17 of the Michigan report*

CDC: Performance Task Design

- Implication
  - The inclusion of Performance Tasks serves an important signaling function: that we value complex, interconnected knowledge and skills
    - How do we balance this important signaling function with testing time?
    - How do we address “the problem of one”?
    - How do we assure that form to form and year to year changes in performance tasks are possible to account for in high-stakes decisions (the equating problem)?

Questions?

- Did the bias or item/task type sections challenge your thinking?
- Do you challenge our thinking on bias and item/task type?

- Any questions, comments, or discussion?

Critical Design Considerations

- Flowing from the information we desire to convey in reports and data files is the need for information from the test (and items/tasks)

CDC: Information Needs

- Implications
  - The more information we want to report, the more data points we need to create that information
    - More reporting categories means we need more items/tasks and more testing time
    - We need adequate information to support each reporting category
  - We need to maximize the data we can extract from our items/tasks
    - We don't want to jeopardize our claims by falling into the trap of creating more information with simpler item types because it is quicker for students to complete
    - We need to extract as much information as possible out of complex tasks (holistic vs. analytical vs. hybrid scoring)
    - We need to offset the cost of scoring complex items/tasks with Artificial Intelligence to the degree supportable
  - We need to prioritize where we need the greatest information
    - Fixed-form (paper and pencil or computer based) testing has a conundrum:
      - For achievement scores, information is most critical at cut points with consequences
      - For growth scores, information across the entire range of scores is important
    - Computer adaptive (target at each student's achievement level)
    - We need to be careful to avoid jeopardizing our claims when prioritizing information (e.g., we shouldn't eliminate parts of the blueprint to meet targeted information needs)

Critical Design Considerations

- Another critical design consideration is the test administration mode (paper and pencil vs. computer-delivered vs. computer adaptive)

- *Addressed to some degree on pages 10-13, 34-37, and 40-41 of the Michigan report*

CDC: Test Administration Mode

- An extremely brief primer on two relatively new modes of administration:
  - Computer-Based
    - Delivery of a fixed-form assessment (may include many different forms)
    - Allows for more flexibility of item/task types above paper and pencil
    - Allow for efficiencies in scoring above paper and pencil
  - Computer Adaptive Testing
    - Has the benefits of computer-based testing
    - Delivers a unique test form tailored to the individual student
    - Reduces testing time by achieving adequate information for each student more quickly through tailoring
    - For high-achieving students, presents an atypically difficult test
    - For low-achieving students, presents an atypically easy test
    - Unlike fixed-form testing, a student's scale score…
      - …is not closely related to students' total scores on the items/tasks
      - …is closely related to the average difficulty of the items/tasks taken
    - **All of these unique CAT characteristics are reduced by other Principled Assessment Design constraints such as alignment and test security (sometimes considerably)**

We assessment people call this a "severely constrained CAT"

CDC: Test Administration Mode

- Implications
  - Paper and pencil testing presents considerable logistical and security concerns
    - Shipping, scanning, scoring present opportunities for losing or mis-reading documents
    - Paper handling increases time between testing and reporting
    - Paper materials allow for unauthorized access and security concerns
    - Many test forms to address security concerns increases workload and possibility for error
  - Computer-based testing
    - Getting all schools and students online is a challenge. We need to assure that the broadest possible range of operating systems and devices is supported.
    - The use of a computer should not reduce any student's access to the test content. How can this be achieved?
  - Consistency of the assessment experience
    - Conflicting with the need to broadly support operating systems and devices, stability of the operating systems and devices is a concern. Many schools rely on old operating systems that are no longer supported.
    - Conflicting with the need to broadly support operating systems and devices, the items/tasks need to be displayed in nearly the same way for all students. Different graphical rending systems and different screen sizes may affect student performance.
    - If some take via paper and pencil and some take via computer, it may affect student performance.

Critical Design Considerations

- Another critical design consideration is the test security approach
  - High-stakes is correlated with inappropriate student and educator testing behavior
  - Remedying a security breach can have dramatic implications on a student, classroom, school, district, or state. It is much better to avoid a security breach

- *Addressed to some degree on pages 8-13 and 40-41 of the Michigan report*

CDC: Test Security Approach

- Implications
  - How can security breaches best be prevented for fixed-form testing (paper and pencil or computer-based)?
    - A single day per test section
      - Severely restricts flexibility in scheduling
      - Decreases test participation
      - A backup form and test date alleviates some decrease in test participation
      - An emergency form reduces impact of a contained security breach
    - Many test forms allow for flexibility in scheduling by randomly assigning forms or administering each form only on one or two days
  - How can security breaches best be prevented for computer-adaptive testing?
    - A deep item bank allows for flexibility in scheduling by giving each student a unique test
  - Highly contextualized items/tasks (such as Performance Tasks) tend to be the most memorable tasks, and are likely to be compromised if care is not exercised
    - Do we create many Performance Tasks around a single theme that can be used as the contextualization so that they can be randomly assigned?
    - Do we cycle through Performance Tasks on so that no single Performance Task is presented on more than, say, five days?

Critical Design Considerations

- Another critical design consideration is test timing and flexibility of testing windows
  - The timing of testing and the timely return of aggregate test results are often in conflict
  - Greater flexibility in testing windows is desirable for some reasons and undesirable for others.

CDC: Test Timing and Flexibility of Test Windows

- Implications
  - The closer to the end of the year, the better for the following reasons
    - Students receive the full year of instruction before being tested on that grade's content standards
    - It is very clear that all that grade's content standards should be on the assessment
  - To a point, the further from the end of the year, the better for the following reasons
    - Summertime program and curriculum evaluations requires final data soon after the end of the school year
    - Accountability designations require data soon after the end of the school year to be published before the beginning of the next

CDC: Test Timing and Flexibility of Test Windows

- Implications
  - The more flexible the testing window, the greater local autonomy is preserved in scheduling
    - Spring break may vary by district
    - Community events may conflict
  - The greater the flexibility in testing windows, the less comparable the test scores are
    - A 12-week testing window means 12 weeks less instruction for some students than for others before testing on the same grade's content
    - A 12-week testing window in consecutive years means that for calculating student growth, some students will have had 24 weeks less instructional time between tests than others
    - How can flexibility and comparability be balanced?

Changing Topics

The July 28-29 In-person Meeting in Laramie

Decisions In Advance of July 28-29 Meeting in Laramie

- Two groupings of members is needed:
  - Three groups to provide feedback on the revised draft of Chapter I (*background and uses/characteristics section* vs. *intended outcomes section* vs. *comprehensive assessment system section*)
    - Email us your first and second choices by July 17th.
  - Two groups to focus on state summative assessment versus a comprehensive assessment system
    - Email us your first choice by July 17th.
  - We will do our best to accommodate your desired assignments if we hear from you by July 17th.
  - We need to achieve balance of roles in the groups, so you may not get your choice.
  - We will email your assignments with pre-readings so you can focus your preparation.

Decisions In Advance of July 28-29 Meeting in Laramie

- Thinking about feasibility for a deliverable and whether the recommendations will be likely to be accepted (we need to design a high-quality Chevy, not a Ferrari)
- If a comprehensive assessment system is proposed, we need to limit the recommendations. We propose the following as a limited set of considerations for the July 28-29 meeting
  - Classroom formative
  - District interim/summative
  - State summative
- What does a district assessment have to do to meet the full information needs for instruction, evaluation, and accountability?
- Does everything need to come from on-demand assessment?

Thank You

- We particularly thank you for your attention to a mostly didactic presentation on design considerations in state summative assessment, which we felt was important for the July 28-29 work in Laramie.

- Any final questions?