



# High-Level Model for an Assessment of Common Standards

---

March 1, 2010  
Stephen Lazer





# This presentation

- › Overview and disclaimers
- › Description of process
- › Design questions, recommended answers, implications of answers
- › Discussion and next steps



# Overview and disclaimers

- › Attempting to do something odd and possibly dangerous
  - › Design an assessment where nobody yet has made final selections among key priorities and where there are no final domain targets.
- › Purpose here is to get us talking
  - › Some of the decisions and choices may be wrong.
- › Any test design is a compromise
  - › One test cannot be all things.
  - › If other people tell you they can . . .



# General solution parameters

- › Common core standards will be adopted by large numbers of states within 1-2 years.
- › The common core standards will cohere across grades so that assessments of the standards will support meaningful estimates of student growth.
- › States and/or consortia may want to measure some attributes beyond those covered in the common core standards.
- › Universal computer-based testing will be possible in 3-4 years (special-needs accommodations may still be given on paper).
- › However, this paper does not assume enough technology availability to be able to test all students in mass administrations.



# General solution parameters

- › Major elements of the new assessment system must be made operational within 3-4 years.
  - › However, this initial operational roll-out does not need to represent the end state system; more elements may be added at later dates.
- › Efficiencies from pooled test development and psychometric work will make possible modest (though not major) increases in the per pupil operational costs of assessment, allowing some use of human scoring in the system.
- › The goals of the common core assessment can only be met by an assessment system, and not by a single test.



# Assumptions about goals (1)

- › Solid and thorough measurement of constructs as defined in the standards, including complex skills
- › While focus on measurement of new higher standards, some underlying need for solid information about children across the ability spectrum
- › Need for accountability as well as instructionally actionable data
  - › Accountability data must meet professional technical standards for rigor
- › Information on the common standards that is strictly comparable across states, districts, individuals, etc.
- › Need to allow for state/consortium augmentation of common standards test



# Assumptions about goals (2)

- › Desire for innovation
- › Desire for aggressive use of technology
- › Desire for system that has positive instructional wash-back
- › Desire for system that can play a role in professional development
- › Want rapid scores and an affordable system (may run contrary to other goals)
- › Want to support teacher/school evaluations
- › **OBVIOUSLY ALL OF THESE GOALS CANNOT BE OPTIMIZED**



# Assumptions about implementation

- › Elements of the system can be staged in over time.
- › K-8 elements may fit one model, while high school elements may fit another.
- › System may cost a bit more than the current system to implement, but increases need to be modest.



1. Are we designing a summative assessment, a formative assessment, or both?

- › **Recommendation: General design should consider or be consistent with an integrated system in which summative and formative components cohere as an information provision system.**



# 1. Why an integrated system?

- › People will need both accountability and instructionally actionable data, and no single test will provide both.
- › Need to reduce pressure on summative tests to provide formative data.
- › If designed well, summative tests could be providers of data to formative systems (perhaps across years).



# 1. Possible pitfalls of the integrated systems approach

- › RFPs/federal funding are likely to cover only summative elements.
- › Formative assessment is as much about professional development and process as about instrumentation.



# NOTE

- › Slides below discuss summative components



## 2. Defining a summative system (K-8)

- › In K-8, there will be end-of-year tests in ELA and math at each grade.
  - › Schools remain organized by grade, and accountability systems like “check-ups” at fixed points in time.
- › However, the final summative system may not be limited to end-of-year tests.
  - › Data gathered over the course of the year, either through periodic standardized assessments, projects, or portfolios, could well become part of the summative system.
  - › Likelihood of this approach to succeed may depend on degree of curricular uniformity that emerges from common standards movement.
  - › Much to be gained; depending on approach, many problems to be solved.



## 2. Defining a summative system (high school)

- › Two possible models: end-of-course (EOC) or end-of-domain (EOD).
- › Either model has advantages, although the former is better suited to ideas like periodic assessment.
- › It is not inconceivable the system would make use of both.
- › More on this later.



### 3. What general design should the assessment have?

- › **Recommendation: Multiple Testing Components**
  - › **Component 1: A test that measures the common core standards, producing scores on these components**
  - › **Component 2: Test materials built to state-specific or consortium-specific standards, to be used in conjunction with common materials; scores based on analyses of common core plus state-specific materials (these materials will not produce scores on their own)**



### 3. Reasons for multiple-component approach

- › In theory, states will adopt all common standards and may augment with 15% of their own.
- › This approach allows solid and unambiguous comparisons across states on the common standards components.
- › Also allows states to augment the common test, work with their own vendors, etc. (less monopolistic).
- › Also allows states not to need to augment materials.
- › Allowing for intermixing of common and state-specific content would be operationally and psychometrically untenable.



## 3. Possible problems

- › States may not like the section-based approach and may want to interweave content.
  - › Is likely not workable in this model.
- › Because of the perceived importance of comparability of scores, does presume a common test.



## 4. What grades and subjects?

- › **Recommendation:**
  - › **Math and ELA (Reading and Writing) end-of-year tests at grades 3-8**
  - › **Math and ELA (Reading and Writing, and possibly Listening and Speaking) EOD in high school or EOC in some selected array of subjects (likely to include Algebra II and Grade 11 English, but these would not be the only ones and in math would not cover the standards)**



## 4. Why these grades and subjects?

- › States used to this data pattern
- › Easiest support for growth modeling in K-8
- › Best interactions with formative components of the system



## 5. Common or state-specific scales; international benchmarking

- › **Recommendation:**
  - › **Single scale and performance levels for common components**
    - › **Possible normative data**
    - › **Possible international benchmarking**
  - › **Separate state-by-state analyses of conjoined sets of state-specific plus common materials**
- › **Why?**
  - › **Allows comparisons on common standards**
  - › **Allows state-specific data**



## 5. Possible problems

- › Some states may want to have their state-specific materials placed on the common scale, which may or may not be problematic (depends on the nature of the materials).
- › Need to decide on an international benchmarking approach.
- › May prove difficult for states to explain different passing rates than shown on the common core components.



## 6. Computer or paper or both

- › **Recommendation: Even though availability of technology could cause complications, we recommend that major elements of the program be computer-based.**
  - › **Paper used for accommodations**



## 6. Why computer?

- › Allows for use of item types needed to measure construct as described in the draft college readiness standards.
- › Allows for aggressive use of electronic scoring.
- › Maintaining comparable computer and paper systems would limit what we could measure and introduce ongoing psychometric issues.
- › Test will not be fully operational for five years, and technology access is likely to grow.
- › Allows for adaptive administration, which we view as a major plus given challenging standards.
- › RFP is likely to demand innovation, which will be limited in paper test.
- › Supports faster score turnaround.
- › Gives a fairly easy way for state or consortium components to be added.



## 6. Implications and possible problems

- › Insufficient availability of equipment could render plan infeasible (or at least affects the initial consortium of states).
  - › Some reviewers and work-group participants viewed the risk as too large.
  - › It might limit which states could initially be in the consortium.
- › Will lead to testing windows of likely 4-6 weeks, which creates security issues, placing upward pressure on item pools.



## 7. What sorts of tasks?

- › **Recommendation: While we will discuss the sorts of items we expect to use, our proposed solution should include an Evidence Centered Design (ECD) process to design task models and test models.**
- › Of course, we do know a good deal:
  - › Standards call for students to be able to use digital media and data, so assessments will need to measure that.
  - › Measurement of emerging standards will require aggressive use of items beyond multiple choice; part of this attachment is because of broader effects on educational system.
  - › We will still need individually reliable scores, so there are some real limits on the use of long items.



## 7. What sorts of tasks?

- › EOY tests will likely have mix of selected response, short answer, and extended answer.
- › Will use scenario approaches as appropriate.
  - › Scenario approaches are defined as sets of problems built around themes and/or contextual areas.
- › Will be developed through ECD process and will be based on cognitive models and not solely evident alignment with standard, which is likely to be insufficient.
- › Some items will make use of digital tools, as skills demand.
- › Examples will be included in Appendix to paper.



## 7. Tasks: possible issues and problems

- › All tasks will likely not be amenable to automated scoring.
  - › Extensive use of human scoring at this scale may be prohibitively expensive and will slow score turnaround, and introduces complexities for pre-equating and adaptive administration (may be manageable).
  - › However, construct is likely to demand use of some human-scored items.
  - › Constructed-response scoring also is viewed by some as important professional development.
  - › There may be ways to control costs, but these do not solve all other problems.



## 7. Tasks: possible issues and problems

- › Need to develop an operational definition of the ELA construct:
  - › How to measure writing and at which levels
  - › What to do about speaking and listening
- › Answers to these questions could drive us away from a single common core system, affecting the comparability of data, which is a key goal



## 8. Adaptive or linear administration?

- › **Recommendation: The common core assessment component should make use of adaptive administration.**
- › Why?
  - › Shorter testing times reduces demand for computer seat-time.
  - › Allows for measurement of higher standards while still providing information about what lower performers know and can do.
  - › Could let us make more effective use of open-ended questions.
- › **Note: By “adaptive administration,” we do not necessarily mean item-level CAT.**



## 8. Issues and decisions

- › NCLB has in the past been hostile to certain elements of adaptive testing we might want to use (off-grade material).
- › Getting an adaptive system started is complex and involves calibration of a large stock of items (for security reasons).
  - › One realistic possibility is that the program would not be able to be fully adaptive in initial year of roll-out.
- › There are a number of methods of adaptive administration; selecting the right one will depend on decisions made elsewhere (like on human scoring).
  - › Likelihood of human-scored items will probably push us to some sort of staged system in which components that cannot be scored by automated systems come after adaptation has occurred.



## 9. Constructed-response scoring

- › System needs to assume the use of items that cannot be machine scored.
- › A great deal of interest in involving teachers in scoring, for various reasons.
- › Need scoring systems in which teachers can participate, but in which scores are checked and moderated to ensure data comparability.
  - › Likely to involve use of distributed scoring systems
- › This may need to change if and when periodic elements enter system, when local scoring might have some real advantages (and problems).



## 9. Scoring (cont.)

- › If goal of teacher involvement is professional development, we must make sure that professional development is thought through (otherwise work will be drudgery).



# 10. Extending the system beyond EOY tests

- › Summative/accountability data does not need to come solely from EOY tests.
- › EOY data could be augmented by periodic standardized tests over the course of the year, or by scores from standardized student projects.
- › This represents a major rethinking of the way summative/accountability data is obtained.



## 10. Periodic assessment elements: advantages

- › Get better data about students
- › Provide some instructionally actionable data from summative system
- › Allow the sort of testing that would not fit well in EOY test
- › Encourages better integration of instruction and assessment
- › Under simple approaches, may be possible to do needed combinations of scores



## 10. Periodic assessment elements: disadvantages/issues

- › Implies some curricular uniformity, at least in terms of aggregation of content and possibly in terms of sequence.
  - › If not, then statistical methods of equating components and combining data from periodic elements must be invented.
- › Even with uniformity, simply more components to keep track of, equate, etc.
- › Decisions needs to be made on how much the components are supposed to stand alone.
- › Local choice on project-based elements will have implications for data comparability.
- › Unless the periodic elements are instructionally valuable, they may appear to be simply “more testing.”



## 10. Periodic assessment elements

- › Even though the issues will need to be carefully considered, we still believe that this approach should be pursued.
- › It may or may not be possible to roll these out as part of the accountability system during the initial roll-out.
- › However, upside is substantial and well worth pursuing.



# 11. Validity and benchmark data

- › Need for validity argument and a related action theory
  - › Most current state programs and NAEP have limited external validity data.
  - › This is likely to not be acceptable here.
  - › In the high-school components, where the construct is college readiness, this is especially true.
    - › Data should support that claim, possibly beyond simple statement that they were built to benchmarked readiness standards.
    - › We should also realize that these data need to be revisited.



# 11. Validity and benchmark data

- › International benchmark data
  - › Be clear about what we hope to gain from these links and benchmarks
  - › How to do benchmarking?
    - › Ensure that content standards are benchmarked
    - › Ensure that performance standards are judgmentally benchmarked
    - › Do some statistical linkages (may not work well, and problems in grades and subjects where there is no linking study)
  - › How often to do (linkages are temporal)



# Thoughts on HS models

- › End-of-domain and end-of-course
- › Choice is largely policy and not technical
- › One could make an argument to use both sorts of assessments



# End-of-domain model

## Advantages

- › Need to maintain only two tests
- › Possible for within HS growth measurement
- › Direct measurement of standards and CCR
- › Easy comparisons across places

## Disadvantages

- › Lacks direct tie to curriculum/instruction
- › Does not support feedback to or evaluation of teachers
- › Not consistent with periodic assessment or standardized project approaches



# End-of-course model

## Advantages

- › Promotes rigorous course-taking and instruction
- › Links assessment and instruction
- › Consistent with project-based and periodic assessment approaches
- › What is done successfully in other countries

## Disadvantages

- › Need to build and maintain several exams
- › Comparability of data for students in different courses limited
- › Teachers and students are not all in core courses
- › Implies real curricular uniformity in courses
- › Growth measurement difficult



# Thoughts on HS

- › In either the EOD or EOC model, test will involve computer delivery
- › Mix of item types is likely to be similar
- › Different approaches are optimized for different goals
- › Policy decision as to which – one might even want to argue in favor of both



# Closing thoughts: What doesn't this model do?

- › Not cheapest possible model
- › Not fastest possible score turnaround
- › Does not claim to provide very extensive formative data out of the summative system



# Closing thoughts: Why use this model?

- › Forward-looking and solid measurement of emerging domains
- › Aggressive but realistic in the period in question – pushes limits but does not assume wholesale invention
- › Takes an evolutionary approach to adding periodic elements
- › Definitely addresses a number of limitations of current summative systems