Considerations for the NY State Assessment System

Jennifer Dunn & Scott F. Marion

National Center for the Improvement of Educational Assessment

New York State Board of Regents

Center for Assessment

June 12, 2017

Assessment in New York

- The Regents have directed SED staff and technical advisors to think through issues and opportunities associated with making changes for the Next Generation state testing system

- We will be discussing:
  - Design considerations and tradeoffs associated with assessment design

Common uses of assessments

- Student Level
  - Measure achievement
  - Measure strengths and weaknesses
  - Make individual student decisions
- School Level
  - Accountability
  - Educator evaluation
  - Program evaluation
- District & State Level
  - Accountability
  - Program evaluation
  - Comparisons

The challenge of assessment design

We want an assessment that can:

- Provide information useful for evaluating programs and interventions
- Provide information for improving teaching and learning
- Provide high-quality data for fair accountability
- Be administered in less than 2 hours
- Be administered during the last week of school
- Deliver results at least a month before school gets out
- Be inexpensive

You can't have it all!

NY Assessment Priorities

- **Reporting Goals:**
  - Student Level
    - Overall Achievement
    - Diagnostic Tool
    - Growth
  - School Level
    - Status
    - Improvement
    - Growth
- **Measurement Goals:**
  - Valued by educators
  - Meets technical quality criteria
  - High proportions of extended response items
  - NY educators involvement in test development
  - Local scoring

Test Design

- **Test design**, like most engineering activities, is a case of **optimization** under **constraints**
- Design Considerations
  - Reporting requirements (student vs. school level)
    - Reliability
    - Subscores
  - Measurement Requirements
    - Content coverage (depth and breadth)
    - Amount of extended-response tasks
  - Also might need to consider certain **design alternatives**:
    - Embedded Field Testing
    - Matrix designs
    - Connection to other assessments

Reporting Requirements

Subscores

- Reduce or eliminate reporting student subscores
  - Advantage: Encourages use of the most valid reported scores (total)
  - Disadvantage: Educators (and some parents) want more than just a total math score, for example, after students have spent several hours taking a test. Note: The ESSA Think Tank recommended retaining subscores.
- Consider School/District Subscores: Use items that are spiraled across students to report subscores at the school or district level.
  - Advantage: Results in reliable and valid subscores at the school level
  - Disadvantage: Subscores are not reported at the student level

Adjust Reporting Requirements

Test reliability considerations

- **Optimize the test reliability given the purpose of the test**
    - **Student Level**
    - **School Level**

Measurement Requirements

Content Representation

- Depth and breadth of content coverage
  - Sample standards across years.
  - Advantage: Allows the measurement of all standards across years. Encourages educators to teach beyond the test.
  - Disadvantage: Not all students would be measured on all standards each year. Makes it harder for educators to predict what will be on the test.

Measurement Requirements

Item types

- Reduce the number of open response items
  - Advantage: Can optimize content coverage while minimizing testing time
  - Disadvantage: May reduce the ability of the assessments to measure complex skills and may send "signals" that unintentionally lower curriculum and instruction expectations
- Increase the number of items/passage
  - Advantage: Highly efficient use of testing time
  - Disadvantage: Tends to be more difficult to develop and field test. May increase costs.

Test Design

Field testing

- Consider embedded field testing for constructed response tasks in lieu of stand alone field testing.
  - Items that need to be tested for future use are administered as part of the operational assessment
  - Advantage: on average will likely shorten overall testing time and lead to a higher quality field test.
  - Disadvantage: would make the operational test longer and complicates the ability to continue localized scoring.

Matrix Sampling

- Matrix sampling involves distributing the full set of test items among multiple forms:
  - Students take only one form
  - All forms are administered at the class or school level

- Advantage: Efficient use of testing time while generating reliable scores at the school (or class) level
- Disadvantage: Students do not take the same items. Does not allow for raw (number correct) scores.

- Hybrids between common and matrix designs (e.g., 50% of the items are common) offer benefits of both designs.

Test Design

Connection to other assessments

- Interim assessments could be designed to measure the same learning targets and use similar types of questions (e.g., performance tasks)
  - **Intended to create coherence between the interim and summative systems**
  - **Modular** assessment designs are tied to specific aspects of the full content standards, but each assessment focuses on just a limited subset of the full domain. These present some logistical challenges from an educator, administrator, and reporting perspective.
- Shift the emphasis of some content/measures **from the summative test to local assessment**
    - Could assess some knowledge and skills in greater depth, but shorten the testing experience

Turn and talk

1. What are some of the most important considerations for you with a new state summative test aligned with the Next Generation Standards?
2. What elements are least important to you? (You must select at least one!)
   a. Reporting subscores
   b. Reliability Considerations (student/school)
   c. Content coverage on state summative test
   d. Use of performance or other open-ended tasks
   e. Stand alone field testing
   f. Expectation that all students would take the same items (e.g., allow for matrix-sampling designs)
   g. Use of a single summative assessment (as opposed to one that was connected to interim assessments)

The importance of stability

- One of the most common uses of assessments is related to monitoring achievement over time
  - Trend Lines
- Any change to the assessment can potentially impact the ability to maintain achievement trend lines
  - Administration policies
  - Content standards
  - Test length
  - Test composition

An **Example** of a Test Development Sequence

Summer 2017

- **Adopt** new and/or modified **content standards**

17-18

- Professional development and curriculum alignment with new content standards
- **Educator/stakeholder involvement in conceptualizing the new assessment design**
- Determine the new assessment design

18-19

- Engage in item development, repurposing existing items and writing new items
- Engage in item tryouts and cognitive laboratories
- Educator/stakeholder involvement (e.g., item review, bias review, data review)

19-20

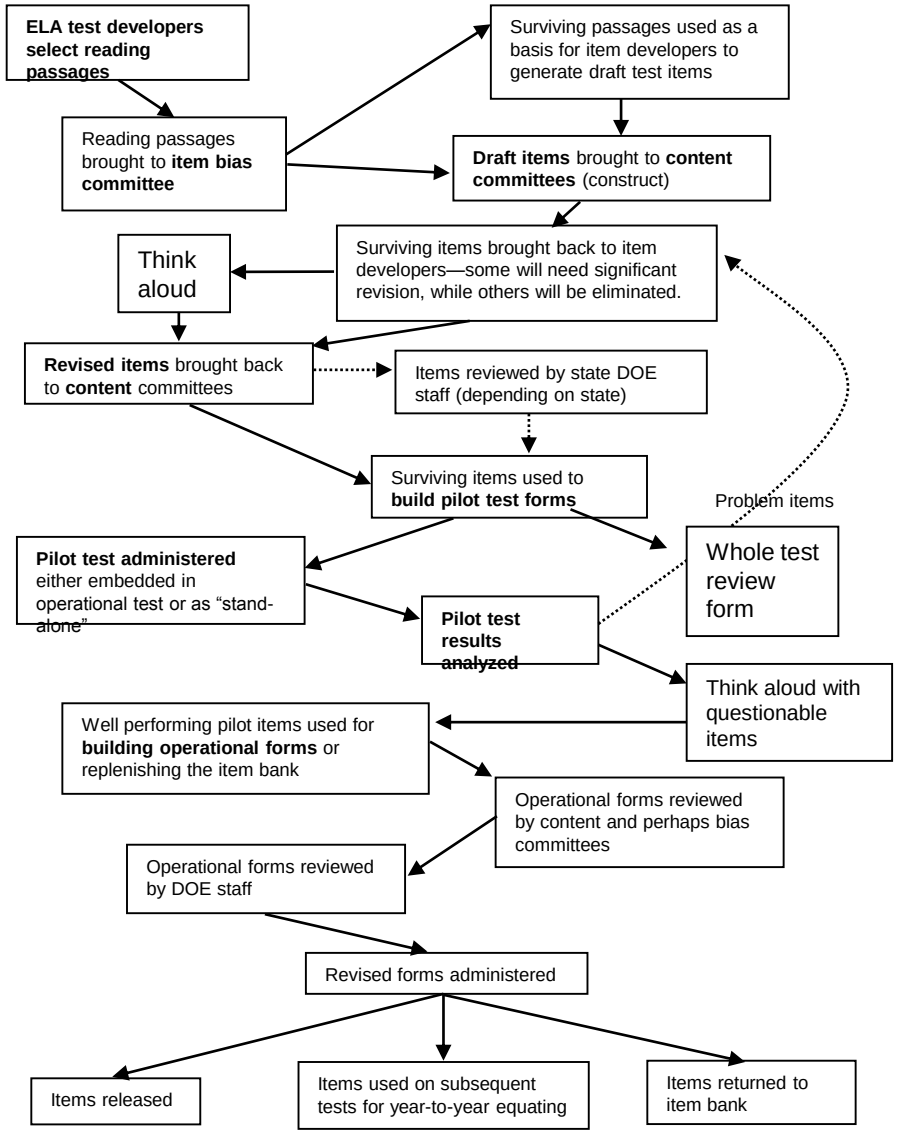
- Continue item development, start building field test "forms"
- Embed field test items in legacy test

20-21

- Review field test results build operational "forms"
- First administration of new operational test
- Set standards for reporting

The Life Cycle of an ELA Test item

This schematic illustrates the many steps involved in developing a test item for an operational test form – once the test design has been finalized. Believe it or not, this is actually a bit of an oversimplification.



How to move forward to a plan…

- Assessment is highly **sensitive** and **visible**
- Broad-based surveys help gather stakeholder opinions, but it is often necessary to turn to a deliberative body to wrestle with the difficult choices (optimization under constraints)
- Many states have turned to ad hoc committees (e.g., Assessment Task Force) to advise policy makers
  - Includes various types of educators from different types of school systems, higher education, business, parents, and others
  - NY may be able to re-purpose and re-configure some of the current stakeholder groups including the technical advisory committee (TAC)

Costs and benefits

- As mentioned earlier, every potential solution carries certain costs

- We need to lay out the obvious tradeoffs as well as consider the potential **unintended negative consequences**

- Again, it is critical **to create a multi-year plan** so that educators and others have **predictable information**

Turn and talk

1. What are some of the key features that you'd like to see as part of a future test design (e.g., performance-based tasks, projects, computer-adaptive, curriculum-embedded assessments)?
2. How important is it for you that the items are developed by NY teachers in future?
3. How important is it that the trend lines are maintained?

Innovative Assessment and Accountability

- Allows for a pilot for **up to seven (7) states** to use **competency-based or other innovative assessment approaches** for use in making accountability determinations
- Initial demonstration period of **three (3) years** with a two (2) year extension based on satisfactory report from the Director of the Institute for Education Sciences (IES), plus another potential two (2) years at the discretion of the Secretary
- **Rigorous assessment, participation, and reporting requirements**
- Subject to a **peer review** process
- May be used with a subset of districts based on strict "**guardrails**," with a **plan to move statewide by end of extension**

Assessment Flexibility Under the Pilot

- **Assessments are not Required to be the Same Statewide**
  - Approved states would have the flexibility to pilot the assessment system with a subset of districts before scaling the system statewide by the end of the Demonstration Authority.
- **Assessments may Consist Entirely of Performance Tasks**
  - Approved states would have the flexibility to design an assessment or system of assessments that consists of all performance tasks, portfolios, or extended learning tasks.
- **Assessments may be Administered When Students Are Ready**
  - Approved states can assess students when they are ready to demonstrate mastery of standards and competencies as applicable.

“Guardrails” for the Pilot

- **Assessment Quality**
  - The state needs to demonstrate that the system of assessments is comprised of high quality assessments that support the calculation of valid, reliable, and comparable annual determinations as well as provide useful information to relevant stakeholders about what students know and can do relative to the learning targets.
- **Comparability**
  - The state needs to demonstrate that its innovative assessment system produces yearly, student-level annual determinations that are comparable across LEAs and to the federally required statewide assessments and for each subgroup of students as compared to the results for such students on federally required state assessments.
- **Scale Statewide**
  - If the state is proposing to administer the innovative assessment system initially in a subset of LEAs, the state must have a logical plan to scale up the innovative assessment system statewide in the State’s proposed demonstration authority period.
- **Demographic Diversity & Similarity**
  - The state can describe how the inclusion of additional LEAs will help the state make progress toward achieving high-quality and consistent implementation across demographically diverse LEAs.

Recapping the small group discussion

Four Regents participated in the "Innovative Pilot" small group at the March 27th meeting and discussed:

- NY should continue to investigate the ways in which NY might take advantage of the flexibility offered in the pilot
- The decision must be "vision driven" and we must be clear about what we hope to accomplish with this pilot
- There was an interest in "starting small" by focusing first on either writing and/or science
- Critical there is a recognition of funding and other resource issues associated with engaging in such a pilot

Questions, comments, discussion

- What are your thoughts on including a sketch of a potential Demonstration Authority application in the State Plan?
- What remaining questions do you have about changing the assessment prior to 2020 or so?
- Other comments and questions?