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

April 4, 2017

Assessment in NY

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

- We will be discussing:
  - Design considerations and tradeoffs
  - Implications of changing the measures

Center for Assessment

Common uses of assessments

- ## Student Level
  - Measure Achievement
  - Measure strengths and weakness
  - Make individual student decisions

- ## School Level
  - Accountability
  - Teacher evaluation
  - Program evaluation

- ## District & State Level
  - Accountability
  - Program Evaluation
  - Comparisons

Center for Assessment

The challenge of assessment design

We want an assessment that:

- Provides information useful for evaluating programs and interventions

- Provides information for improving teaching and learning

- Provides high-quality data for fair accountability

- Is administered during the last week of school

- Can deliver results at least a month before school gets out

- Is inexpensive

Pick one!

Center for Assessment

NY Assessment Priorities

- **Reporting Goals:**
  - Student Level
    - Overall Achievement
    - Diagnostic Achievement
    - Growth
  - School Level
    - Status
    - Improvement
    - Growth

- **Measurement Goals:**
  - Valued by Educators
  - High proportions of extended response items
  - Local Development
  - Local Scoring

Center for Assessment

Adjust Reporting Requirements

Subscores

- Reduce or eliminate reporting student subscores
  - <u>Disadvantage</u>: Educators (and perhaps parents) want more than just a total math score, for example, after students have spent several hours taking a test. Note: The 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

Test reliability

- Reduce test reliability by shortening the test.
- <u>Disadvantage</u>: Student scores will be less reliable. Reduces the capability of the assessment to measure student growth.

Center for Assessment

Reduce the Measurement Requirements

Content Representation

- Reduce depth and breadth of content coverage
  - Sample standards across years
  - <u>Disadvantage</u>: Not all students would be measured on all standards each year.

Item types

- Reduce the number of open response items
  - <u>Disadvantage</u>: May reduce the ability of the assessments to measure complex skills and reduce educator buy in
- Increase the number of items/passage
  - <u>Disadvantage</u>: Tends to be more difficult to develop and field test. May increase costs

Field testing

- Consider embedded field testing in lieu of stand alone field testing.
  - Items that need to be tested for future use are administered as part of the operational assessment
  - <u>Disadvantage</u>: would make the operational test longer and might have an impact on localized scoring
  - <u>Advantage</u>: will shorten overall testing time and will lead to a higher quality field test

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

- <u>Advantage</u>: Efficient use of testing time while generating reliable scores at the school (or class) level

- <u>Disadvantage</u>: Students do not take the same items. Does not allow for student scores

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

Center for Assessment

Connection to other assessments

- Interim assessments could be designed to measure the same learning targets and using 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
- Shift 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?

2. What elements are least important to you (you must select something)?

   a. Reporting subscores

   b. Student-level reliability (impacts measurement of growth)

   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)

Center for Assessment

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 valid achievement trend lines
  - Administration policies
  - Content standards
  - Test length
  - Test composition

If you want to measure change…

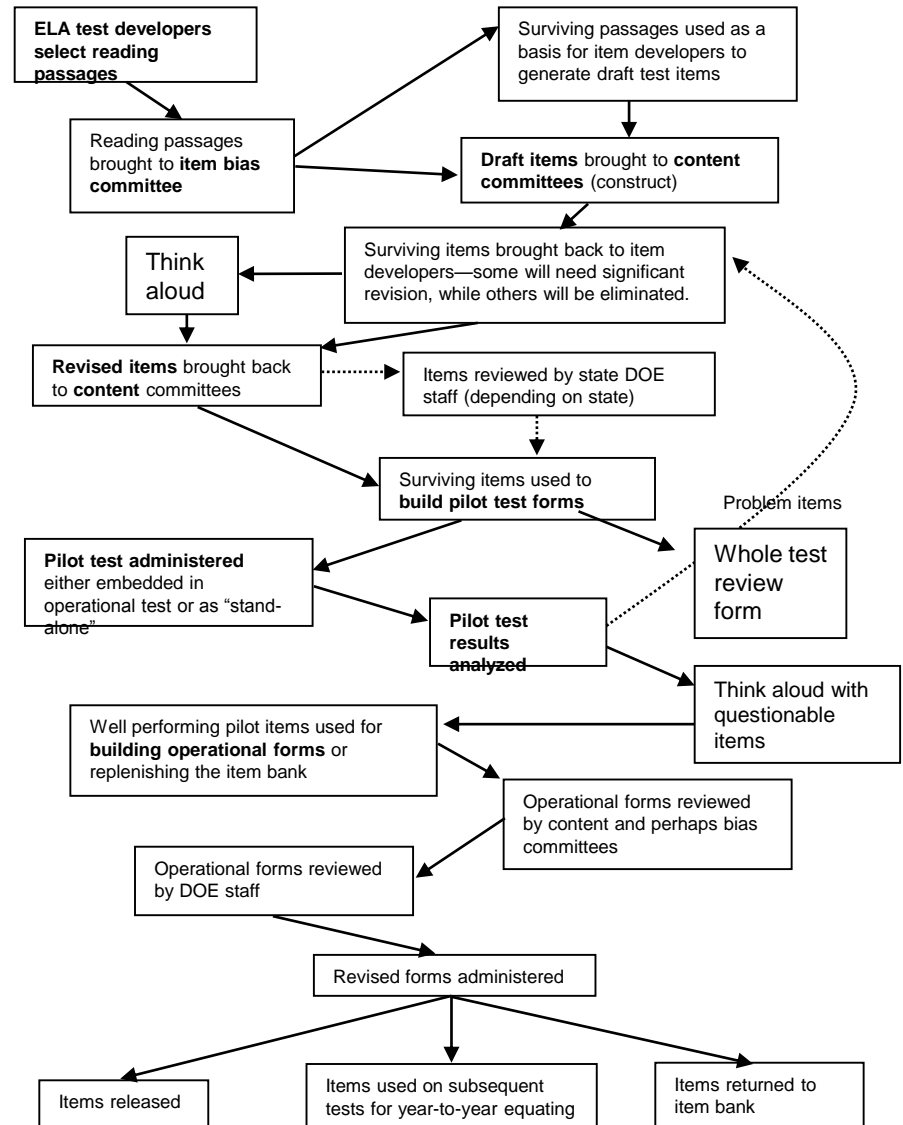# don't change the measure

Center for Assessment

Why does this take so long?

- Scott created the following graphic to illustrate the various steps involved in developing items for a large-scale, standards-based assessment...

Center for Assessment

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. Believe it or not, this is actually a bit of an oversimplification.

ELA test developers **select reading passages**

Reading passages brought to **item bias committee**

Surviving passages used as a basis for item developers to generate draft test items

Draft items brought to **content committees** (construct)

Surviving items brought back to item developers—some will need significant revision, while others will be eliminated.

Think aloud

Revised items brought back to **content** committees

Items reviewed by state DOE staff (depending on state)

Surviving items used to **build pilot test forms**

Problem items

Whole test review form

Pilot test administered either embedded in operational test or as "stand-alone"

Pilot test results analyzed

Think aloud with questionable items

Well performing pilot items used for **building operational forms** or replenishing the item bank

Operational forms reviewed by content and perhaps bias committees

Operational forms reviewed by DOE staff

Revised forms administered

Items released

Items used on subsequent tests for year-to-year equating

Items returned to item bank

Center for Assessment

The Importance of Stability

- We are beginning to understand some of the **policy challenges** facing the Regents and SED

- We recommend the Regents **minimize the number of changes** in the assessment system prior to the necessary change to measure the new standards

- At a minimum, we need **to create a clear 5+ year plan** to provide **predictable information** as we move into our new accountability system

Center for Assessment

How to move forward to a plan…

- Assessment is highly **political** 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, politics, parents, and others

  - For example, see this report from Wyoming that was used to guide the recent RFP.

Costs and benefits

- As I mentioned earlier, every potential solution carries certain costs

- We need to layout the obvious tradeoffs as well as considering the potential **unintended negative consequences**

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

Center for Assessment

Turn and talk

1. How important is it for you that the trend lines are maintained?

2. How important is it for you that the items are developed by NY teachers?

3. 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)?

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 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

- Maybe used with a subset of districts based on strict "**guardrails**," with a **plan to move statewide by end of extension**

Center for Assessment

Innovative Assessment and Accountability

An ***Innovative Assessment System*** *means a system of assessments that may include:*

(1) competency-based assessments, instructionally embedded assessments, interim assessments, cumulative year end assessments, or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments;

(2) assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs.

Center for Assessment

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.

Latest on the Demonstration Authority

- Final Rules were published on December 8, 2016, which means we are passed the 60 day window to employ the Congressional Review Act (CRA)

- The Secretary **MAY** release an application for states

- We have not heard much talk about such an application, but this could be due to the lack of high-level staff in place at USED

Center for Assessment

Recapping last week's 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

- There was a recognition of funding and other resource issues associated with engaging in such a pilot

Therefore, the small group recommended including NY's intention to apply for the Demonstration Authority as part of the State Plan

Center for Assessment