The Changing Landscape of Statewide Assessment: Shifts towards Systems of Assessments

Virtual Session in NCME Fall Conference
August 24, 2020
Session

On the Shift Towards Balanced Assessment Systems: Past, Present and Future  Brian Gong, Center for Assessment

Developing a Validity Research Agenda for Louisiana’s Innovative Assessment Demonstration Authority Pilot  Nathan Dadey, Center for Assessment; Michelle Boyer, Center for Assessment

On the Opportunity Provided in Creating an Innovative Assessment: Design Considerations and Agile Test Development in an Innovative Pilot  Abby Javurek, NWEA; Paul Nichols, NWEA; Garron Gianopulos, NWEA

Discussant & QA Moderator: Carla Evans, Center for Assessment

Questions & Comments: Put in Chat/QA
On the Shift Towards Balanced Assessment Systems: Past, Present and Future

Brian Gong, Center for Assessment

Overview

• Background: Problem definition and tools
• Systems of assessment
• Some innovative visions involving interim assessments
• A note about innovation
Problem definition and tools

Need for more, better, more timely assessment information

Theory of action and assessment validation
We’re in a new generation of assessment policy and design

Title 1 Evaluation, ESEA, 1965

State assessments & accountability since 1980’s: performance assessments, Writing, Soc. Studies, etc.

Push for different theories of action and assessment design

• Recognition that accountability’s summative assessments do not provide enough information to directly inform improved learning, at the right time, under the appropriate governance (control)
  • Need validation: starting with specific interpretive/use arguments
  • Validity research agenda

• Also often coupled with calls for different accountability system(s) under different theories of action
Assessments should be set within a larger theory of action

• The test interpretation usually informs some reasoning about additional information, and action. Most learning consequences are results of that reasoning and action, not directly of the test interpretation
  • Diagnosis [analysis of test results] → Prescription [what to do] → Treatment [do]

• Program evaluation provides models for how to evaluate claims (ToA) about relationships between diagnosis/actions and outcomes
• Validation of assessments through interpretive/use and validation arguments
Examples: Different uses imply different assessments

**Improvement** theories of action (examples) and associated needed test information

• Schools that perform relatively very low on state tests should be identified annually by the state for support
  • Assessments of annual stable performance of current year content are comparable across schools, time

• Districts/Schools should focus on improving core instructional effectiveness for all students
  • There should be assessments useful for informing within-cycle instruction and assessments for informing program evaluation closely tied to curriculum, instruction, conditions of school/district

**Instructional** theories of action (examples) and associated needed test information

• Present then remediate
  • Assess current content after instruction; grain-size: within-unit remediation

• Remediate before current unit
  • Assess previous year content before instruction

• Differentiate to keep on-grade
  • Assess current unit content and key (few) pre-requisites before instruction
Systems of assessments with a focus on interim

Assessment systems within larger systems

Vertical/horizontal coherence
Assessment systems

• Multiple assessments
• System(s) of assessment by design
  • Coherence
  • Comprehensiveness
  • Continuity
  • Utility

Assessment systems are parts of larger systems

- Accountability, instruction, curriculum, equity policy, educational funding
- Larger systems’ goals, resources, constraints shape the assessment system
  - Examples:
    - System improvement (e.g., external accountability vs. internal formative evaluation)
    - Instructional approaches (e.g., Instruct-Remediate; Remediate-Instruct; Instruct on-grade with Differentiation)
- Assessment system should be coherent with larger system
  - Valid interpretations provide valuable information that can be used
  - Timely
  - Efficient / appropriate resources
  - Right governance / who owns the assessment information
  - Adaptable to changes in larger system
Focus on “Interim”

• Time: between a beginning and end
• Partial: one piece of a series or set
• Formative – summative evaluation
• Not confirmed, “acting”

• For this paper: Has a scale score intended to be generalized beyond the particular test items, combined and compared (Perie et al., 2009)
Systems of Interim Assessments

- Multiple assessments during one year (horizontal)
  - “Interim,” “modular”

- Interim assessments in lieu of summative (vertical)
  - IADA – ESSA “Innovative Assessment Demonstration Authority”
    - Produce summative information about student proficiency in relation to state content and performance standards of sufficient quality to use in state accountability system (comparability)
  - Efficient? – If already administering interims, could eliminate summative?
Elements of claims for interim assessments that are especially important and challenging

- **Vertical coherence** for interim assessments that are designed/used as a part of a larger assessment system (how interim assessments relate to summative and formative assessments) – and where assessment fits in learning system

- **Horizontal coherence** for interim assessments that are designed as a set, and to interim assessments outside the set (how interim assessments relate to each other)
  - Is there a structure (e.g., boundary, sequence) to the content (e.g., grade level)

- Claims related to **time** (when the claim applies)
  - Often involves whether at another time there would be learning/forgetting; may involve assumptions about instructional supports or other context

- Claims/scores that embody **aggregation** of evidence across assessments
  - Claims when there is contradictory evidence across assessments
Examples of vertical coherence across types of assessments

• Monitoring of functioning at different levels of system for that level’s purposes (formative: student learning, interim: district program improvement, summative: state policy; national/international policy)
  • Governance assessment ecosystems: different assessments provide different information for different users—owned by different entities and therefore usually loosely coupled and barely coordinated
  • For example, international (TIMSS), national (NAEP), state, district, school, classroom, student

• Drill-down diagnosis: More general test → more specific test → even more specific test to identify specific weakness and reasons for it

• Support and preparation for valued outcome: formative: inform micro learning/teaching; interim: feedback on performance in less structured, more independent, larger performance contexts; summative: performance of record on target assessment

• Progressive attainment of increasing complexity and expertise
• Learning supports for progressive attainment
• Periodic external demonstration
**Examples of horizontal coherence across interim assessments**

**Assessment target:** growth or **progress over a course of instruction** (each implies a different test design)

- “Looking forward/Looking back” – timing relation to instructional use
- **Construct definition over time**
  - Opportunities to show **more accurate** performance on the same content
  - Divide up content domain into (sequence of) **different content**
  - **Increasing independence**/less scaffolding in solving similar problems
  - **Increasingly sophisticated** ways to solving the same/similar problems
  - Solving **more complex** problems
  - Application of **self-evaluation** to improve

Challenges for design and claims of interim assessments in systems of assessment

Challenges in design and possible solution approaches
Coherence across set of interim assessments (within year)

• Use – little coherence vs. high coherence
• What construct is—what develops over time
• What claim is
• How content is organized (what is assessed when)
• How performance at a point in time is viewed as evidence
• How assessments are compared with each other
• How performance at multiple points in time are viewed as evidence; if aggregated, how
  • Scaling
Challenges to claims when purposes are combined/shifted: example

<table>
<thead>
<tr>
<th>Claims, intended information/interpretation, and design</th>
<th>Summative/policy/programmatic</th>
<th>Interim</th>
<th>Formative/instructional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalization to broad domain</td>
<td></td>
<td></td>
<td>Often as specific a content or subskill as possible</td>
</tr>
<tr>
<td>Student can perform independently</td>
<td></td>
<td></td>
<td>Student can learn interactively with teacher, peers, resources</td>
</tr>
<tr>
<td>Stable at the end of year or after</td>
<td></td>
<td></td>
<td>At that moment (it should change)</td>
</tr>
</tbody>
</table>
Some innovative visions involving interim assessments
Characteristics of some innovative projects involving interim assessments

• Increase cognitive complexity of assessment demands, e.g., performance assessments
• Relate more strongly to specific curricula
• Embed interim assessments into curriculum, i.e., local choices about administration, use locally (e.g., grades) as well as summative; sometimes local scoring
• Use multiple interim assessments to provide summative score in lieu of summative test
• Integrate interim assessments into vertical system (e.g., integrated content specifications, scale, selection and administration supports)
More innovations

• Specify content to link claims to item development (e.g., range ALDs)
• Develop items to ALDs to enable front-end (embedded) alignment and standard-setting
• Develop processes and tools (e.g., principled assessment design) that build for validity and utility
• Develop culture and systems that allow for more rapid development try-outs, feedback, and improvements (e.g., continuous improvement, agile, scrum), particularly for new and not-yet-routinized projects
Using theories of action, program evaluations, and assessment validation to clarify aims and possible benefits of interim assessments
Evaluation to improve

• Evaluation of **goals** and intended **outcomes** – social policy evaluation and construction

• Evaluation of **theory of action** and associated **programs** – program evaluation; formative program evaluation especially useful to those enacting the theory of action

• Evaluation of **assessment** quality – validation arguments and evidence
  • Different theory, approaches, criteria?
    • Sufficiency,
    • Standardization
Evaluation of assessment in terms of validity and usefulness

- **Learning ToA**
  - Student learning model
  - Instructional model

- **Assessment ToA**
  - Claims (tiered)
  - Assessment design

- **Claims validation**
  - Analytic evidence
  - Empirical evidence
A note about innovation
We’re in a new generation of assessment policy and design
We’re in a new generation of assessment policy and design

Federal policy leading vs. lagging driver of reform

Divergent Theories of Action (and Goals?)

Accountability State Custom

State Custom On-line

Interim in lieu of State

Assessment systems

Instructional Assessment Systems

Instructional Assessment Systems - COVID

Instructional

Instructional Assessment

NRT
Developing a Validity Research Agenda for Louisiana’s Innovative Assessment Demonstration Authority Pilot

Nathan Dadey & Michelle Boyer

The National Center for the Improvement of Educational Assessment
Outline

1. Context & Design
2. Program Theory: Theory of Action & Claims
3. Scaling
4. Conclusions & Next Steps
1. Context

Enabling factors for Louisiana's Innovative Assessment Demonstration Authority Pilot
State Assessment as a Continuation of Reform

- **2013**: ELA Guidebooks 1.0 released. ELA Guidebooks framework and text sets developed.
- **2014**: ELA Guidebooks 2.0 piloted with 10 districts.
- **2015**: ELA Guidebooks 2.0 released statewide.
- **2016**: Successful application to IADA Pilot & assessment development in grade 7.
- **2018**: Partial grade 7 administration.
Guidebooks 2.0:

- Open source curriculum
- Designed to meet Louisiana's criteria for high quality instructional materials
- Unit based, with each unit:
  - organized around a central idea with one or more corresponding texts
  - containing daily lessons, classroom assessments, instructional guides, writing samples and more
- Use is voluntary, but the majority of schools have adopted them

www.nciea.org
A System of Assessments Perspective: The Status Quo

**State.** Large-Scale Standardized Accountability Assessment

**District.** Interim/Benchmark Assessments

**Classroom.** Quizzes & Tests

Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4
A System of Assessments Perspective

State. Large-Scale Standardized Accountability Assessment

District. Interim/Benchmark Assessments

Classroom. Quizzes & Tests

Guidebooks 2.0

# A System of Assessments Perspective

<table>
<thead>
<tr>
<th>State</th>
<th>District Selected Unit</th>
<th>District Selected Unit</th>
<th>Common Unit</th>
</tr>
</thead>
</table>

- **State**:
  - Aug.
  - Sept.
  - Oct.
  - Nov.
  - Dec.
  - Jan.
  - Feb.
  - March
  - April
  - May

- **District**:
  - Aug.
  - Sept.
  - Oct.
  - Nov.
  - Dec.
  - Jan.
  - Feb.
  - March
  - April
  - May

- **Classroom**:
  - Aug.
  - Sept.
  - Oct.
  - Nov.
  - Dec.
  - Jan.
  - Feb.
  - March
  - April
  - May

- **District Selected Unit**
  - Aug.
  - Sept.
  - Oct.
  - Nov.
  - Dec.
  - Jan.
  - Feb.
  - March
  - April
  - May

- **Common Unit**
  - Aug.
  - Sept.
  - Oct.
  - Nov.
  - Dec.
  - Jan.
  - Feb.
  - March
  - April
  - May
Each End-of-Unit Assessment:

- Is meant to assess students’ ability to understand and to build knowledge from the unit texts, and express that knowledge and understanding in writing.
- Follows the same general blueprint.
- Is administered in two sessions, each lasting an hour.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Window 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Students take one of two unit assessments.</td>
<td>The Giver</td>
<td>Written in Bone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Grade 7 ELA Design
Grade 7 ELA Design

Window 1
- Students take one of two unit assessments.
- The Giver
- Written in Bone
- End-of-Unit Report #1

Window 2
- Students take one of four unit assessments.
- The Giver
- Written in Bone
- A Christmas Carol
- Memoir
- End-of-Unit Report #2

Window 3
- Students take both a unit assessment and writing task.
- Behind the Scenes
- End of Year Writing Task
- End-of-Year Report

Timeline:
- Aug.
- Sept.
- Oct.
- Nov.
- Dec.
- Jan.
- Feb.
- March
- April
- May

Fall
- Spring
Program Theory

Theory of Action & Supporting Interpretive Argument
Need for a Theory of Action

Making good on the opportunities provided by the ELA 2.0 Guidebooks and the IADA waiver requires a theory of action that connects program inputs to ultimate outcomes.

Ideally, once articulated, the theory of action then helps define the needed score interpretation(s) (e.g., Bennett, Kane & Bridgeman, 2011).

→ Although we present our work here as linear, development was anything but. We have been continually revising and iterating.
“Validity is the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.”

(Emphasis Added, AERA, APA, NCME, 2014)
Theory of Action

• Developing a detailed theory of action lead us to differentiate between within year and end of year inputs, action mechanisms and outcomes.
  ▪ E.g., smaller theories of action within the larger theory of action\(^1\)

• These smaller theories contain specific use cases and supporting score interpretations.

A set, or better yet a system, of assessments can – potentially – support multiple claims and associated use cases.

\(^1\)E.g., what Forte & Hebbler (2004) call a “grand theory of action”.
Educators and leaders receive professional development at the Teacher Leader Summit.

Leaders and educators identify classrooms and schools in need of support.

Leaders provide direct support to teachers in identified classrooms and schools.

Educators receive professional development in the first of four ongoing workshops.

Educators interpret results in light of prior instructional practice and Guidebook instructional guidance.

Educators make plans to improve the fidelity of future instruction using Guidebook resources.

Educators implement Guidebook instruction with greater fidelity, drawing on the guidebook practices, including those outlined in the diverse learners cycle.

Improved Student Learning Throughout Unit 1

Part of Overall School Performance Score

School Identification of CSI And TSI
End of Year Working Logic Model

High-level Summary:
• Assessment results signal the need for instruction based on the Guidebooks to be implemented with fidelity
• Educators, support from local leaders and state experts, will implement Guidebooks with greater fidelity
• Resulting in improved student learning in the next year

Use Cases:
• Signal the need to focus on guidebook instruction
• Support state systems of accountability

Educators receive professional development at the Teacher Leader Summit
Leaders provide direct support to teachers in identified classrooms and schools
Educators implement Guidebook instruction with greater fidelity, drawing on the guidebook practices, including those outlined in the diverse learners cycle
Educators interpret results in light of prior instructional guidance
Educators make plans to improve the fidelity of the instruction
Leaders provide resources and support to teachers in identified classrooms and schools
Educators receive professional development in the first of four ongoing workshops

Input
Action Mechanism
Effect

High-level Summary:
• Assessment results signal the need for instruction based on the Guidebooks to be implemented with fidelity
• Educators, support from local leaders and state experts, will implement Guidebooks with greater fidelity
• Resulting in improved student learning in the next year

Use Cases:
• Signal the need to focus on guidebook instruction
• Support state systems of accountability
“Validity is the degree to which evidence and theory support the **interpretations of test scores** for proposed uses of tests.”

(Emphasis Added, AERA, APA, NCME, 2014)
First Claim

Use Cases:
• Signal the need to focus on guidebook instruction.
• Support state systems of accountability.

Claim 1:
Students can apply their knowledge and skills gained from the units of instruction to read and write effectively, and to generate meanings from texts.
Within Year *Working* Logic Model

**Input**
- Assessment End-of-Unit 1 Results
- Assessment Interpretive Materials

**Action Mechanism**
- Educators implement Guidebook instruction with greater fidelity, drawing on the guidebook practices, including those outlined in the *diverse learners cycle*.
- Educators adjust Unit 2 instruction based on assessment results and information from their classroom practice.

**Effect**
- Improved Student Learning Throughout Unit 2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Spring</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within Year *Working* Logic Model

High-level Summary:
- In addition to motivating greater fidelity of guidebook instruction,
- Educators **adjust instruction in the subsequent unit** based on the results from the prior unit, by drawing on the instructional practices outlined in the guidebooks and supports
- Resulting in improved student learning in the next unit

Use Cases:
- Signal the need to focus on guidebook instruction
- Guide instruction in subsequent units


**Fall**
**Spring**
Second Claim based on The System

Use Cases:
• Signal the need to focus on guidebook instruction
• Guide instruction in subsequent units

Claim 2:
Students can apply their knowledge from a unit to:
- make sense of the texts from that unit
- make sense of texts related to that unit
- Write effectively about these unit and unit related texts
Restated

Support multiple claims and associated use cases within a system involves:

- Partitioning student performance by use case (and even designing assessments, or portions of assessments to satisfy one of the claims and uses)
- Using multiple measurement models to produce information based on subsets of the data to support the intended interpretations and uses
3. Scaling in Support of Claim 1

Producing a “single summative score”
On A Summative Score

• The production of a single summative score based on multiple assessments has been an area of interest and limited research (e.g., Wise, 2011, Dadey & Gong, 2017)
  ▪ Rooted in both measurement and value judgments

• Claim 1 requires the production of such a score:

Students can apply their knowledge and skills gained from the units of instruction to read and write effectively, and to generate meanings from texts.
On A Summative Score

• Current approach is to **pool item responses across windows**, treating the entire set of item responses as if they are one larger assessment
  ▪ Doing so means that any estimates of item difficulty reflect difficulty right after learning occurs, instead of at the end of the year
Window 1
Students take one of the two unit assessments.
- The Giver (Form A)
- Written in Bone (Form A)

Window 2
Students take one of the four unit assessments.
- The Giver (Form B)
- Memoir
- A Christmas Carol
- Written in Bone (Form B)

Window 3
Students take both a unit assessment and writing task
- Behind the Scenes
- End of Year Essay

Window 1
- The Giver (Form A) I1 ... I10
- Written in Bone (Form A) I1 ... I10

Window 2
- The Giver (Form B) I1 ... I10
- Written in Bone (Form B) I1 ... I10
- Memoir I1 ... I10
- A Christmas Carol I1 ... I10

Window 3
- Behind the Scenes I1 ... I10
- End of Year Essay I1 I2
Refining Claim 1

Students can apply their knowledge and skills gained from the units of instruction to read and write effectively, and to generate meanings from texts.

• Collapsing data across windows means that data is aggregated in a “grade book” manner, meaning performance in each of the three windows is equally accounted for in the total score
  ▪ Barring differences in item discrimination
Refining Claim 1

Students can apply their knowledge and skills gained from the units of instruction to read and write effectively, and to generate meanings from texts.

• Whether this approached results in scores that are comparable enough to the statewide assessment remains an open question
  ▪ There are multiple options for weighting performance across the assessments
Grade 7 ELA Administration 19-20

Window 1
Students take one of the two unit assessments.

The Giver (Form A)
Written in Bone (Form A)

Window 2
Students take one of the four unit assessments.

The Giver (Form B)
Memoir
Written in Bone (Form B)
A Christmas Carol

End of Year Essay
Written in Bone (Form B)

Behind the Scenes

Common Items
Grade 7 ELA
Administration 19-20

Student 1
Student 2
Conclusions and Next Steps
Future Directions

• Engage in formative evaluation of the program as a whole, and in doing so collect specific validity evidence for the interpretive and validity argument
  ▪ Based on the five sources of validity evidence

• Continue to iterate on the program in areas that can be changed (e.g., use and interpretation of end-of-unit scores)
On the Opportunity Provided in Creating an Innovative Assessment: Design Considerations and Agile Test Development in an Innovative Pilot

NWEA
Working Presentation, NCME
August 24, 2020
Abby Javurek, Paul Nichols & Garron Gianopulos
Assumptions about standard assessment practice challenged as part of IADA projects

COVID-19 disruptions further challenged assumptions about the learning ecosystem

Transdisciplinary approaches that borrow from traditional assessment approaches, systems thinking, and software development allow for Agile Development of assessments that better meet customer needs

Allows for multiple Theories of Action for different users and buyers, centered in their role in accomplishing the Job-to-be-done (JTBD)
The Approach: A Transdisciplinary Process

+ Back to First Principles
  - Testing and replacing out assumptions (continuously checking and refining)
  - Hypothesize, test, revise cycles
+ Systems Thinking
  - System expansion to the extended learning ecosystem
+ Agile Development
  - Personas
  - Job to Be Done
  - Short cycles/ sprints and A/B testing
Back to First Principles: New Assumptions

- Solutions should prioritize time for classroom instruction throughout the year.
- Solutions should prioritize data that helps teachers inform instructional activities.
- Solutions should build across the year and not be a “post-mortem”.
- Solutions should take into account prior information about students.
- Measuring the process of learning and tracking how students are growing is a priority.
- Scoring and data reporting can happen in near real-time.
Understanding the system to be changed:
The Learning Diamond (Nichols & Ferrara, 2014)

- **School Ecosystem**: Who are the players that are shaping the ecosystem?
- **Professional Learning**: What is planned, what is needed?
- **Instruction**: Lever to create change
- **Curriculum**: Maintenance of local control as a priority
- **Theory of Learning**: Rooted in Range Achievement Level descriptors
- **Assessment**: Better interim data, and making a separate summative redundant
- **Home Ecosystem**: More important than ever and we are still understanding it
Who are the Key Personas and what are the Jobs to be Done (JTBD)?

• State and district leaders; teachers and school leaders; parents and students

• Accelerate Learning for all students

• Close Achievement Gaps

• Foster Readiness for rigorous high school college and career readiness coursework
Agile Test Development

+ Borrowing from the software industry:
  - Flexibility
  - Rapid Cycles of refinement
  - Constant value testing to prioritize focus
+ Agile method during test design
  - Vision
  - Design Sprint
  - Hypothesis testing
    + Technically feasible?
    + Add value?
+ Fail early, not late
+ Solution must fit within customer/state/national policy constraints
What this means:

<table>
<thead>
<tr>
<th>Prioritize This</th>
<th>Above this</th>
<th>In practice:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals and Interactions</td>
<td>Process and Tools</td>
<td>cross functional teams</td>
</tr>
<tr>
<td>Working Prototypes</td>
<td>Excessive Documentation</td>
<td>Purposeful documentation re: validity and efficacy; not check lists</td>
</tr>
<tr>
<td>Customer collaboration</td>
<td>Rigid Contracts</td>
<td>Customers embedded in design and development processes</td>
</tr>
<tr>
<td>Responding to Change</td>
<td>Following a plan</td>
<td>Treating this like an optimization problem with small bets and A/B testing; testing assumptions</td>
</tr>
</tbody>
</table>
From Traditional Design and Development

Adapted from Downing’s 12 Steps for effective test development (2006)
To Agile Cycles

- Define Success Criteria
- Preliminary Specifications
- A/B testing
- Specification validation
- Information Gathering and Validation
- Documenting Validity

- Item Design and Development
- Production and Administration
- Standard Setting
- Reporting
Process to Arrive at the Best Field Test Plan and Test Design

1. Define success criteria*
2. Define preliminary test specifications
3. Define administration constraints*
4. Document technological platform capabilities
5. Delimit adaptive test design options
6. Define item pool needed
7. Narrow field test options*
8. Vet field test options with decision makers and stakeholders*
9. Converge on best plans and design after multiple iterations

*Customer collaboration and consensus needed
Design and Development Challenges examined during rapid cycle testing: A balancing act

- Level of detail vs test length
- Accountability vs utility
- Validity studies vs political realities of time
- Instructional relevance vs comparability
- Field testing and comparability
- Customization vs standardization
Innovative solutions offer potential value but also bring potential risk.

- Current SOPs should only be replaced with an innovative solution if the value add is substantial.
  - Research is necessary to collect evidence of feasibility (low risk) and utility (value) of an idea.

\[
\text{Customer Value} = \sum (\text{Importance} \times \text{Relevance})_n - \sum (\text{Cost} \times \text{Importance})_c.
\]

Where \( n \) is customer needs and \( c \) is customer costs.
Implementation

Innovation driven by policy changes due to concerns about:

- Timeliness of results
- Actionable data
- Disconnects from instruction (assessments feel like events, not like learning)
- Over testing/Testing time
- Making stakes too high

In Georgia SB 362: Up to 10 districts or consortiums of districts to apply to pilot innovative assessments aligned with content standards; Sets up clear roles for State Board Education, Georgia DOE, Office of Student Achievement, Independent third-party evaluator to examine comparability.
Make “Small Bets” at Key Decision Points

Each hypothesis is a small bet; If any bet succeeds, value will be added via increased adaptivity and measurement sensitivity. All desired outcomes in the TOA depend on measurement sensitivity.
GMAP: Through-year assessment summary

• Fall, winter, and spring assessments will measure student performance relative to the state blueprint while also adapting as needed.

• Each assessment will measure growth and grade-level performance.

• Summative scores will be generated at year’s end.

• The result – richer data about student performance during the school year, a consistent testing experience built on a single “source of truth,” and the elimination of an extra summative test in the spring.
• Durable parts to accomplish JTBD
  • Multiple tests
  • Adaptivity
  • Through year growth
  • Instructionally useful data
“Iceberg” problem: Unfinished learning accumulates and persists, hindering the ability for many students to become ready for college and career.

Solutions being tested...

Source: https://www.icebergproblem.org/
Nuancing Designs to help meet JTBD: Range Achievement Level Descriptors (RALDs)

- RALDs explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated for a standard across achievement levels for each standard and achievement level on the assessment.

- RALDs provide intended content-based interpretations of what scale scores within a particular achievement level represent. Teachers can use RALDs to determine how students with different scale scores within the different achievement levels may differ in their abilities.

- Based in learning science, validated with local educators, RALDs drive item and test development.
Going Forward: Studies for Interpretation

- Stakeholder reporting needs: via surveys and focus groups using representative samples of educators and families. One study outcome will be newly Stakeholder reporting needs: Method -surveys using representative samples of prioritized reporting lists aligned to the JTBDs.

- Rapid prototyping of reports: to iteratively refine reports to support the JTBD of each stakeholder, at the level appropriate for that stakeholder.

- Validation studies: to examine the degree to which intended score representations and recommendations result in accurate and useful feedback (is the data being used and not just valid for the interpretation?), and the degree to which instructional shifts and learning gains are realized.
Contact Us:

- Abby.Javurek@nwea.org
- Paul.Nichols@nwea.org
- Garron.Gianopulos@nwea.org
Discussant Remarks & Panel Questions

Carla Evans

Center for Assessment

NCME Session: The Changing Landscape of Statewide Assessment—Shifts Towards Systems of Assessments

August 24, 2020
### Themes Across Presentations

#### Shifts Towards Systems of Assessments Related to Statewide Assessment

<table>
<thead>
<tr>
<th>Key Barriers</th>
<th>Two Sides of the Same Coin?</th>
<th>Key Facilitators</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Federal requirements and constraints (e.g., comparability, summative score, validity and reliability) for IADA programs</td>
<td></td>
<td>• Robust theories of action</td>
</tr>
<tr>
<td>• Communication among various stakeholders and levels (state, district, school, classroom, etc.)</td>
<td></td>
<td>• Agile development</td>
</tr>
<tr>
<td>• Implementation fidelity—need window into the classroom to gather information about TOA</td>
<td></td>
<td>• Buy-in from teachers</td>
</tr>
<tr>
<td>• Research and evaluation agendas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Political pressures and changes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[www.niea.org](http://www.niea.org)
Questions

1. What elements of a theory of action need to be in place in order to improve student learning? Are these elements consistent across assessment systems, or can they vary?
   • How does the currently implemented program reflect this?

2. Changing instruction tends to be a recurrent theme with respect to innovative assessment systems.
   • What does change to instruction mean? What are key features of improving instruction?
   • Do you think the outcome measure in the currently implemented program is sensitive enough to pick up on changes to instruction and student learning? Explain why or why not.
Questions, Cont’d

3. To what extent is PD an integral part of any TOA intended to improve student learning? Who should be the focus of PD and why? What are the challenges with such an approach to date?

4. What are some unexpected events that have happened during program development and what has allowed you to learn from unexpected occurrences (e.g., COVID)?

5. What is the role of flexibility in these systems (e.g., what aspects of the system can be changed to fit local contexts, and what cannot)?