Through-Year Assessment Virtual Convening
November 15-16, 2021

The National Center for the Improvement of Educational Assessment
Monday, November 15, 1:00-2:30 PM ET
Defining terms, considering aims, and diving into key design features.

Monday, November 15, 3:00-5:00 PM ET
Connecting use cases and claims, and the designs that support them, together to consider needed evidence.

Tuesday, November 16, 1:00-2:30 PM ET
In depth consideration of key big picture technical and logistical issues.

Tuesday, November 16, 3:00-5:00 PM ET
What will it take to make through-year assessment systems work to support students and educators?
Session 2: Claims, Designs, and Evidence
Through-Year Virtual Convening, November 15, 2021

Brian Gong, Nathan Dadey, Will Lorie & Scott Marion
The National Center for the Improvement of Educational Assessment
Invited Panelists

Leslie Nabors-Olah  
ETS

Courtney Bell  
WCER, UW

Karen Barton  
NWEA
1. **Framing**
   Considering the logic and evidence of our theories of action and interpretive arguments

2. **Use Cases & Claims**
   - “Additional” use cases & claims
   - Summative use cases & claims

3. **Panelists Remarks**
   In depth considerations from invited participants

4. **Question and Answer Session**
   Facilitated Audience Interaction
1. Framing
A Definition of Through-Year Assessments

The defining characteristics of a through-year\textsuperscript{1} assessment program are that it is:

- Administered through multiple distinct administrations across a school year, and
- It is meant to support both
  - the production and use of a summative determination, and
  - at least one additional aim: instructional, monitoring, logistical

\textsuperscript{1}We suggest the term “through-year” assessment be used as a general term to describe these models, regardless whether the time period is a “year,” “semester,” or “course.”
# Through-Year Assessment Design Challenges: Multiple Purposes

<table>
<thead>
<tr>
<th>Location of Through-Year Assessment Challenges (ToA, IUA, Design, Evidence)</th>
<th>Intended Use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Assessment</strong></td>
<td><strong>Summative</strong></td>
</tr>
<tr>
<td>State Summative</td>
<td>Widely worked on and researched; much technical support</td>
</tr>
<tr>
<td>Interim</td>
<td></td>
</tr>
</tbody>
</table>
## Through-Year Assessment Design Challenges: Multiple Purposes

<table>
<thead>
<tr>
<th>Location of Through-Year Assessment Challenges (ToA, IUA, Design, Evidence)</th>
<th>Intended Use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Assessment</strong></td>
<td><strong>Summative</strong></td>
</tr>
<tr>
<td><strong>State Summative</strong></td>
<td>Widely worked on and researched; much technical support</td>
</tr>
<tr>
<td><strong>Interim</strong></td>
<td>Lots of examples of designs and uses; little formal ToA, IUA; less technical documentation</td>
</tr>
</tbody>
</table>
# Through-Year Assessment Design Challenges: Multiple Purposes

<table>
<thead>
<tr>
<th>Location of Through-Year Assessment Challenges (ToA, IUA, Design, Evidence)</th>
<th>Intended Use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Assessment</strong></td>
<td><strong>Summative</strong></td>
</tr>
<tr>
<td>State Summative</td>
<td>Widely worked on and researched; much technical support</td>
</tr>
<tr>
<td>Interim</td>
<td>Lots of examples of designs and uses; little formal ToA, IUA; less technical documentation</td>
</tr>
</tbody>
</table>
### Location of Through-Year Assessment Challenges (ToA, IUA, Design, Evidence)

<table>
<thead>
<tr>
<th>Type of Assessment</th>
<th>Summative</th>
<th>Instructional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State Summative</strong></td>
<td>Widely worked on and researched; much technical support</td>
<td>Very limited solutions <em>so far</em> (due to timing; limited subscores due to unidimensionality)</td>
</tr>
</tbody>
</table>
| **Interim** | Most through-year assessment focus:  
  - How can traditional summative criteria be met? or  
  - Can new criteria be developed? | Lots of examples of designs and uses; little formal ToA, IUA; less technical documentation |
Why?  
What motivations underlie the shift to a through year assessment program?

How?  
What is the through year program? What are the intended outcomes and how will they be achieved?

What?  
What is the set of assessments and corresponding assessment evidence?

What evidence is needed?
Through-year programs are acting as interventions and should be evaluated as such.
<table>
<thead>
<tr>
<th>Theory of Action</th>
<th>Interpretive Argument</th>
</tr>
</thead>
</table>
| **Intended Design** | • Inputs  
|                    | • Actions  
|                    | • Outcomes |
| **Supporting Evidence** | • Are the actions implemented with fidelity\(^1\)?  
|                    | • Do they lead to the intended outcomes\(^1\)? |
|                   | • Claims  
|                   | • Design Features |
|                   | • Are the claims about what students know and can do supported? |

\(^1\)Drawn from *Bennet, Kane & Bridgeman, 2011*
Building Out Connections

**Theory of Action**

- Are the actions implemented with fidelity?
- Do they lead to the intended outcomes?

**Interpretive Argument**

- Is the assessment information provided **sufficient** for the given actions?
- Do the actions use the assessment information **appropriately**?
- Are the claims about what students know and can do supported?
Use Cases & Claims

**Theory of Action**

**Interpretive Argument**

**Use Case**

- School Identification & Support
- Instructional or Monitoring Use

**Claim**

- Summative Claim
- Additional Claim(s)
Be as explicit as possible in the statement of the theory of action, especially in the statement of action mechanisms and intended effects, so as to allow for meaningful evaluation.

Bennett, Kane & Bridgeman, 2011, p. 46
2. Uses & Claims
2.A Instructional Claims & Uses
Use Cases & Claims

Theory of Action

Use Case

Summative Use

Instructional Use(s)

Claim

Summative Claim

Instructional Claim(s)
What are the specific actions various actors are meant to take in reaction to the program and the information it produces? In particular, how will these actions lead to better classroom teaching and learning?
“Informing instruction” can include many possible actions, e.g., Abrams, McMillan and Wetzel (2015):
- Modifications to whole class instruction
- Working with students in small groups
- Providing individualized support

Even actions at this level of specificity are not detailed enough to support instruction.

Ideally, these actions are informed by research on score use.
Let’s consider an example theory of action for instructional use for a through-year assessment program (i.e., knowing the results also are intended to inform a summative determination) based on administration of an assessment in the fall semester.
Example High Level Theory of Action

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Actions</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretive Guide</td>
<td>Teachers make instructional adjustments</td>
<td>Student Learning Improves</td>
</tr>
<tr>
<td>Assessment Literacy Professional Development</td>
<td>Results are used to inform summative determination</td>
<td></td>
</tr>
</tbody>
</table>

- **Inputs**:
  - Interpretive Guide
  - Assessment Literacy Professional Development

- **Actions**:
  - Teachers make instructional adjustments
  - Teachers learn to better use assessment data

- **Effects**:
  - Student Learning Improves
  - Results are used to inform summative determination

**Example High Level Theory of Action**

|------|------|-----|------|-----|

**Aug.**

**Sept.**

**Oct.**

**Nov.**

**Dec.**
Example High Level Theory of Action

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Actions</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the specific claim and interpretive use argument: What information provided by the assessment will inform teachers’ and students’ actions?</td>
<td>What is the specific theory of action: What actions will teachers and students take to improve student learning?</td>
<td>Student Learning Improves</td>
</tr>
<tr>
<td>How is the assessment designed and implemented to produce that information?</td>
<td>What conditions are required for that to happen, e.g., what do teachers need to know and be able to do to take those actions?</td>
<td>Results are used to inform summative determination</td>
</tr>
</tbody>
</table>

What evidence supports this theory of action?
What evidence supports the validity and usefulness of this interpretation of the assessment information?
Teachers make instructional adjustments

- Teachers access reports in a timely fashion
- Teachers understand the score reports (mastery of math subdomains)
- Teachers triangulate assessment results with classroom assessments
- Teachers determine what kinds of instructional supports are needed
- Teachers implement supports at the individual and small group level

Student misconceptions on subdomain concepts are addressed
Specific Assumptions

- Teachers have access to, and understand, a variety of instructional supports connected to the math subdomains
- Teachers can use assessment evidence to select amongst the various kinds of support
- The assessment evidence is needed to select amongst these kinds of support
Assessment provides needed information.

Supporting Detail

Assessment Scores and Claims about Mastery of Key Math Subdomains so that Student misconceptions on subdomain concepts are addressed.

Math subdomains are important for understanding current content and for learning subsequent content.

Claims of “mastery” lead to appropriate actions (e.g., go on).

Claims of “mastery” incorporate appropriate depth, breadth, cognitive complexity, generalizability, etc.

Claims about “Non-Mastery” help teachers identify student misconceptions on subdomain concepts.

Scores are reliable and precise enough to support claims and actions.

Scores are reported in way and time that support intended interpretation and use.

Student results can be used in summative determination.
Typical Constraints Posed by Summative Use

• Focuses on state content standards only
• Assessment instruments, administration conditions, times during the year, etc. are highly standardized; carefully designed accommodations and modifications provided to promote comparability and fairness
• Carefully controlled item presentation, mode, and test design to allow equating for comparable scores
• Scoring quality controlled for high accuracy and consistency to officially determined criteria
• Security: item exposure controlled; performance is attributable only to student without input from others or unfair advantage
• Underlying construct and scale usually fit with a unidimensional measurement model established at a certain point in time
2.B Summative Claim & Use
Use Cases & Claims

Theory of Action

Use Case

Comparable Annual Determinations

Instructional Use(s)

Claim

Interpretive Argument

Summative Claim

Instructional Claim(s)
Values drive claims. Claims drive our choice of aggregation methods, not the other way around.

Designs that distribute content in multiple modules across the year require complex claims.
To support annual determinations, we need a single summative score.

The creation of a single summative scores involves not only the application of an aggregation method\(^1\), but also consideration of values and corresponding claims.

\(^1\)Here we include both the application of a measurement model as well as additional post hoc steps like taking the maximum score.
Restated, every aggregation method reflects specific values and supports specific claims.
School Identification & Support

Value Judgement(s)
What value is placed on:
• Performance during the year?
• Performance at the end of the year?
• Changes in performance across the year?

Claims
What inference do we want to make about what students know and can do?, e.g.,:
• About “typical” student performance across the year?
• About student performance at the end of the year?

Score Creation
Implementation:
• Is the aggregation done within a measurement model, or in addition to a measurement model?
• How are the models, and thus time, addressed?

Theory on how learning occurs over time.
One way to understand this complexity is through comparisons to classroom grading.

E.g., is your class grade the sum of unit tests? A pass or fail based on a single end of year test?

To start, let’s consider claims based on the end of the year.

See a blog by Lorie (2021) on this topic.
The question is how to support claims about the end of the year using information across the year.
School Identification & Support

Methods that Support End of the Year Claims

“Modular Mini-Summative” Design & End of Year Claims

The question is how to support claims about the end of the year using information across the year.

Content Domain

<table>
<thead>
<tr>
<th>Module 1</th>
<th>Module 2</th>
<th>Module 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>Spring</td>
<td></td>
</tr>
</tbody>
</table>

• Use scores from modules 1 & 2 as priors for estimates of student scores on module 3
• Use an adaptive model that uses scores from modules 1 & 2 as starting points, but with full recovery

These approaches represent our best guess about, and current understanding of, emerging designs. In a series of blogs Gong (2021a and 2021b) considers the tensions in supporting end of year claims.
These kinds of methods have been suggested, but not yet been implemented operationally.

Careful investigation of both summative score creation and learning & forgetting and opportunity to learn should accompany these approaches.

Gianopulos (2019) provides a comprehensive treatment of proposed summative score creation, for both end of year and across year claims.
Take Away Points

1. Values drive claims. Claims drive our choice of aggregation methods, not the other way around.

2. Designs that distribute content in multiple modules across the year require complex claims.
Value Judgements & Claims: End of Year

Increasing Complexity In Supporting End of Year Claims (due to the interaction of content and time)
Here the interpretations are more complex:

- Performance on each content domain near the time of instruction?
- Number of mastered standards when “ready to test”?

This complexity arises because of the interaction of content and timing.
Value Judgements & Claims: Across Year

The interaction of content and time is a large problem of practice.

Additional work is needed to develop claims and supporting aggregation methods.

There some notable across year approaches, e.g., the aggregation model used by DLM.
Values drive claims. Claims drive our choice of aggregation methods, not the other way around.

Designs that distribute content in multiple modules across the year require complex claims.
3. Participant Remarks
What is an “instructional use”?

- **Information**
  - Status and/or growth scores
  - Diagnostic information

- **User**
  - District/school leadership
  - Teacher support
  - Teacher
  - Student

- **Use**
  - Confirm content coverage
  - Identify students
  - Provide additional instruction
  - In-class grouping
  - Compare with other assessment results
  - In-class grouping
  - Respond to student thinking with teaching moves (Jacobs & Empson, 2016)
    - Understanding student sense-making
    - Exploring strategies
    - Encouraging other strategies
    - Connecting thinking to symbolic notation
    - Posing a related problem

Leslie Nabors-Olah
ETS
What is an “instructional use”?

Information

- Status and/or growth scores
  - District/school leadership
  - Teacher support
  - Teacher

- Diagnostic information
  - Teacher support
  - Teacher
  - Student

User

- District/school leadership
- Teacher support
- Teacher

Use

- Confirm content coverage
- Identify students
- Provide additional instruction
- Compare with other assessment results
- In-class grouping

School activities

- Respond to student thinking with teaching moves (Jacobs & Empson, 2016)
  - Understanding student sense-making
  - Exploring strategies
  - Encouraging other strategies
  - Connecting thinking to symbolic notation
  - Posing a related problem

Copyright © 2021 by Educational Testing Service. All rights reserved. ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). 462301852
What is an “instructional use”?

Organizational strategies

Instructional strategies

Goertz, Oláh, & Riggan, 2009
4. Facilitated Q&A
Supplemental Slides
Value Judgements & Claims: End of Year

Structuring the content & administration based on these assumptions

Content Domain
Direct Content Instruction

Achievement
End of Year
### Five Most Frequent Instructional Uses Evident in the Research on Interim Assessments

<table>
<thead>
<tr>
<th>Use Description</th>
<th>$k$</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Broad claim to modify or improve instruction</td>
<td>14</td>
<td>A total of 86% of teachers reported modifying instruction based on interim assessment results (Clune &amp; White, 2008).</td>
</tr>
<tr>
<td>2. Identify students to provide additional support</td>
<td>13</td>
<td>Results were used to identify students for supplemental instruction (e.g., software, working with volunteers, afterschool tutoring; Shepard et al [2011]).</td>
</tr>
<tr>
<td>3. Identify what content to reteach</td>
<td>10</td>
<td>The administration hoped teachers would reteach with new strategies (Bulkley et al., 2010).</td>
</tr>
<tr>
<td>4. Improve score on the summative assessment</td>
<td>10</td>
<td>Guide schoolwide improvement efforts to meet AYP (especially in low-performing schools; Bulkley et al., 2010).</td>
</tr>
<tr>
<td>5. Differentiate instruction</td>
<td>9</td>
<td>Identifying students with similar patterns of performance on the assessment and using that to construct groups to differentiate instruction (Blanc et al., 2010)</td>
</tr>
</tbody>
</table>

*Note.* Table created based on coding a total of 20 studies featuring interim assessments.
Full Consideration of Summative Score Creation
Use Cases & Claims

Theory of Action

Use Case

School Identification & Support

Instructional Use(s)

Claim

Interpretive Argument

Summative Claim

Instructional Claim(s)
To support annual determinations, we need a single summative score.

The creation of a single summative scores involves not only the application of an aggregation method\(^1\), but also consideration of values and corresponding claims.

\(^1\)Here we include both the application of a measurement model as well as additional post hoc steps like taking the maximum score.
Restated, every aggregation method reflects specific values and supports specific claims.
School Identification & Support

Summative Claim

**Value Judgement(s)**

What value is placed on:
- Performance during the year?
- Performance at the end of the year?
- Changes in performance across the year?

**Claims**

What inference do we want to make about what students know and can do?, e.g.,:
- About “typical” student performance across the year?
- About student performance at the end of the year?

**Score Creation**

Implementation:
- Is the aggregation done within a measurement model, or in addition to a measurement model?
- How are the models, and thus time, addressed?

Theory on how learning occurs over time.
One way to understand this complexity is through comparisons to classroom grading.

E.g., is your class grade the sum of unit tests? A pass or fail based on a single end of year test?
“Modular Mini-Summative” Design & End of Year Claims

The question is how to support claims about the end of the year using information across the year.
School Identification & Support

Methods that Support End of the Year Claims

"Modular Mini-Summative" Design & End of Year Claims

The question is how to support claims about the end of the year using information across the year.

Module 1  Module 2  Module 3

Content Domain

Fall  Spring

• Use scores from modules 1 & 2 as priors for estimates of student scores

• Use an adaptive model that uses scores from modules 1 & 2 as starting points, but with full recovery

These approaches represent is our best guess about, and current understanding of, emerging designs. In a series of blogs Gong (2021a and 2021b) considers the tensions in supporting end of year claims.
Value Judgements & Claims: End of Year

These kinds of methods have not yet been implemented operationally.

Careful investigation of both summative score creation and learning & forgetting and opportunity to learn should accompany these approaches.

Gianopulos (2019) provides a comprehensive treatment of proposed summative score creation, for both end of year and across year claims.
Value Judgements & Claims: End of Year

"Modular Mini-Summative" Design
Each module (a) covers the entire content domain and (b) is identical in terms of content coverage

"Modular Standards" Design
Each module covers an individual standard

Increasing Complexity In Supporting End of Year Claims (due to the interaction of content and time)
Value Judgements & Claims: Across the Year

• Alternatively, we could restructure the claims so that they are *not* rooted in the end of the year, e.g., a claim about:
  ▪ “Typical” student performance across the year?
  ▪ “Best” student performance across the year?
• Doing so means explicitly considering the value attached to each module and its associated portion of the content domain.
• Critically, states are currently constrained to report scores relative to the “full set” of grade-level standards.
School Identification & Support

Value Judgements & Claims: Across the Year

- Simple Sum
- Maximum Score (or above some cut point)
- Simple or Weighted Averages
Here the interpretations are more complex:

- Performance on each content domain near the time of instruction?
- Number of mastered standards when “ready to test”?

This complexity arises because of the interaction of content and timing.
Value Judgements & Claims

Supporting a summative claim may follow this flow:

- **Making assumptions** about, or empirically investigating, student learning & forgetting, opportunity to learn, or both.
- Structuring the content & administration based on these assumptions or findings and, or,
- Statistically adjusting for variation in student learning & forgetting and opportunity to learn.

These ideas are drawn primarily from Gianopulos (2019) and secondarily from Wise (2011) and Bennett et al. (2011).
Open Questions for the Participants

• What evidence is needed to support the interpretive argument of the summative claim based on the through-year assessment program?

• Designing for multiple claims and uses entails various trade-offs. What trade-offs do we have to make to support a summative claim, and do any of those trade-offs limit other uses?
  ▪ How do we collect evidence to understand whether other claims and uses are being impacted?