Through-year Assessment

Nathan Dadey, Brian Gong, Yun-Kyung Kim, and Edynn Sato*

*Authors are listed alphabetically. All authors contributed equally to the conceptualization and development of this module.

Authors

Nathan Dadey

Brian Gong

Yun-Kyung Kim

Edynn Sato

About the Authors

Nathan Dadey	Nathan Dadey, Ph.D. is a Senior Associate at the Center for Assessment. His work focuses on the design, scaling, and use of educational assessments, particularly assessments used for accountability purposes. His work addresses issues that threaten the validity of assessment and accountability operational programs. Dr. Dadey uses psychometric and statistical methods to manage practical problems, including issues related to combining interim assessment data, dimensionality of alternative assessments, subscores, and vertical scales. He aims to produce methodological and applied work that contributes to improved understanding and use of assessment results in policy contexts.
Brian Gong	Brian Gong, Ph.D. is co-founder and Senior Associate at the Center for Assessment. He has been involved with creating policies, models, and criteria for promoting validity, reliability, and credibility in both assessments and accountability systems. Dr. Gong has helped develop solutions including educationally valuable and technically defensible state accountability systems and innovative assessments (e.g., science performance, writing portfolio, learning progressions, growth, non-cognitive, comprehensive assessment systems). He was a member of the committee tasked with revising the Standards of Educational and Psychological Testing and was the co-author of content methodology to implement the CCSSO Criteria for Procuring and Evaluating High-Quality Assessments.
Yun-Kyung Kim	Yun-Kyung Kim is a doctoral student at the University of California, Los Angeles in the Social Research Methodology division with an emphasis on advanced quantitative methods. She received a Master's degree in Educational Measurement and Evaluation from Seoul National University. Ms. Kim has been involved in the design and implementation of measurement and evaluation studies, with application of latent variable models including item response theory models. Her research focuses on latent variable models for longitudinal data and inferences on growth, change, and future trajectory, mainly in the context of English language proficiency assessment and admission setting.
Edynn Sato	Edynn Sato, Ph.D. is CEO of Sato Education Consulting LLC and also is Director of Psychometrics & Research for WIDA at the University of Wisconsin-Madison. She is an experienced Peer Reviewer of State Assessments for the U.S. Department of Education, serves on several technical advisory committees, recently was co-Principal Investigator of a grant and lead designer for an alternate English language proficiency assessment, and she has contributed substantively to the development of a number of other large-scale assessments for accountability. Her research focuses on academic English language processes, opportunity structures and cultural responsiveness, and assessment fairness and validity, with particular interest in alternative and innovative ways to measure what multilingual learners and students with disabilities know and can do.

Learning Objectives

1. Define key terminology related to through-year assessment.
2. Discuss underlying concepts and considerations.
3. Become familiar with the current range of purposes and uses, related designs, and key challenges.
4. Become familiar with methods for addressing challenges in support of the assessment's validity and utility.

Module Sections

- Introduction
- Section 1: Context
- Section 2: Major Design Elements
- Section 3: Examples, Key Challenges, and Methods to Address the Challenges
- Section 4: Considerations for Implementation
- Summary

Notes About This Module

- Emerging area of work
  - Currently no consensus in the field or established/standard practices
  - Information included in this module reflects currently available work and work in development as of Summer 2023

- The information presented in this module reflect the thinking and perspectives of the authors.

Module Citation

Dadey, N., Gong. B., Kim, Y., & Sato, E. (2024). Through-year Assessment [Digital ITEMS Module 35]. *Educational Measurement: Issues and Practice, 43*(1),97-98. https://doi.org/10.1111/emip.12595

Context

1

Section Learning Objectives

1 Context

Become familiar with Federal requirements for state assessments for accountability

List desired areas for the improvement and innovation of assessment that have implications for a through-year assessment model

Define “through-year assessment”

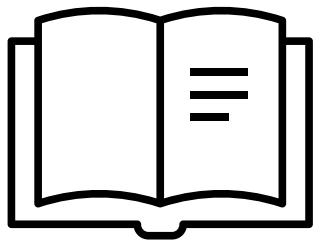
Distinguish through-year assessment from other types of assessment (e.g., interim, formative, summative)

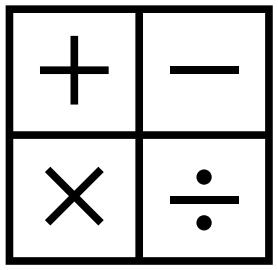
Federal Requirements for State Assessments for Accountability

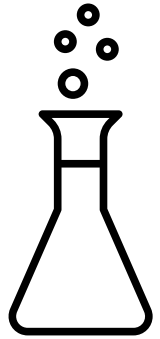
- Every Student Succeeds Act (ESSA) reauthorization of the Elementary and Secondary Education Act (ESEA)
  - Challenging standards
  - Aligned assessments
  - Accountability

Federal Requirements for State Assessments for Accountability

- ESEA requires all states to test all students
  - Annually in **reading or language arts** in grades 3 through 8 and once in high school
  - Annually in **mathematics** in grades 3 through 8 and once in high school
  - In **science** once in each of the following grade spans: 3-5; 6-9; and 10-12







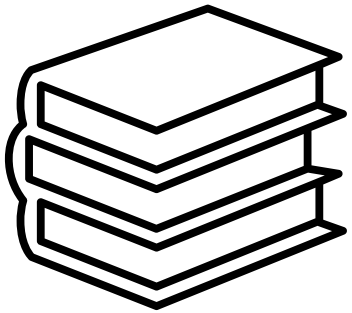
Federal Requirements for State Assessments for Accountability

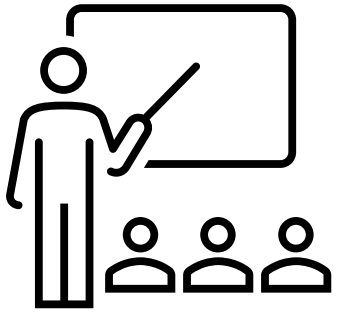
- Of note vis-a-vis through-year assessment:
  - ESSA encouraged what has become known as through-year assessment through a short provision that allowed for a state's accountability assessments to "be administered through multiple statewide interim assessments" (ESSA, §1111(b)(2)(B)(viii)).
  - Innovative Assessment Demonstration Authority (IADA) has been used to support work on through-year assessment

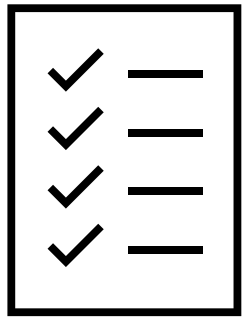
Federal Requirements for State Assessments for Accountability

- For students with disabilities and English learners:
  - Assessments used for accountability purposes (e.g., reading, mathematics, science) must include appropriate accommodations for students with disabilities and English learners.
  - ESEA allows for the use of an **alternate assessment** for students with the most significant cognitive disabilities.
  - ESEA requires all students identified as English learners, including those with the most significant cognitive disabilities, who are enrolled in K-12 schools be assessed annually using the state's **English language proficiency assessment** until exited from English learner services (reclassified).

Desire for Improvement and Innovation of Assessment: General

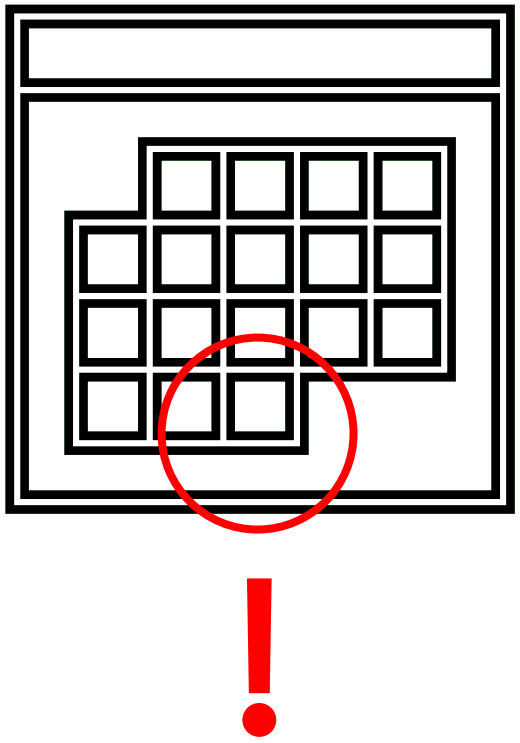






- Gauge how well schools are serving students as educators work to help students recover from instructional and learning disruptions related to the COVID-19 pandemic.
- State initiatives (e.g., related to standards revisions; instructional innovations; assessment life cycle)

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model



Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- Quality, coherence, and integration of information
  - Domain sampling, directness and completeness of measurement of achievement (Koretz, 2008)
  - Coherent, aligned information

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- Granularity of information
  - Descriptive and prescriptive information

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- Use of assessment for learning
  - “Instructionally useful” information
  - Actionable information

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- Timeliness of information

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- Customization/personalization
  - Students learn and demonstrate understanding in a range of ways (Sato, 2023; Sato & Kim, 2022)
  - Culture and cultural diversity can manifest in information having different psychological meaning across different cultural groups (Pearson & Garavaglia, 2003; Sato, 2017; Solano-Flores, 2019)

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- Measurement of growth, change, and progress
  - Arguments for longitudinal analyses to achieve desired quality of test scores (An, Ho, & Davis, 2022)
  - Avoid wrongfully attributing seasonal learning loss (Soland & Thum, 2019)

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model

Desire for:

- More efficient testing
  - Reinforce primacy of instruction and recency of learning
  - Teacher selection of content to reflect instructional scope and sequence (Clark & Karvonen, 2021)
  - Students have multiple opportunities to demonstrate what they know and can do

Desired Areas for the Improvement and Innovation of Assessment: Implications for a Through-year Assessment Model - Summary

Summary

Through-year assessments have the potential to:

- Integrate assessment and instruction
- Yield more descriptive and prescriptive information
- Equip educators with actionable data
- Accommodate the range of different ways students learn
- Adjust to individual student knowledge and skill development trajectories

Definition

There are many different through-year assessment designs, all of which generally meet this general definition (based on Lorié et al., 2021a; Dadey & Gong, 2023):

> A through-year assessment program consists of multiple distinct assessments administered across the school year where information from the multiple assessments is (i) combined to yield a summative determination of student performance to support federally required systems of school identification and support, and (ii) used to support another purpose or purposes (e.g., logistical, administrative, or monitoring).

Definition: Addendum

- Many, in fact, most programs, are choosing only to use the last assessment to create the single summative score.
  - This practice does not align with the definition of through-year assessment presented on the previous slide which requires that information be used from all assessments.
    - This requirement was drawn from the overview information in the Race to the Top Fund Assessment Program which states that a "student's results from through-course summative assessment must be combined to produce the student's total summative assessment score for that academic year" (2010, p. 18, 178).
  - However, the field is generally referring to these programs as through-year.

Overview of Relevant Assessment Models

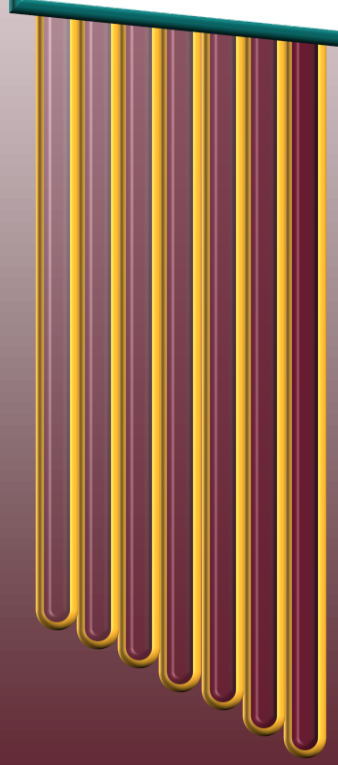
Assessment	Administration	General Purposes and Uses
Through-year	Assessment in multiple parts administered at different times over the course of a school year	➢ Determine whether and to what degree students have learned the material taught ➢ Determine annual learning progress and achievement ➢ Evaluate students' annual progress and achievement across groups and jurisdictions ➢ Evaluate effectiveness of educational programs ➢ Measure progress toward improvement goals ➢ Make course-placement decisions • Focus can be on short-, intermediate- and long-term learning goals and objectives • Inform instruction • More extensive assessment of constructs (e.g., fine-grained learning trajectories throughout the year)
Summative	Typically, administered at the end of a project, unit, course, semester, program, or school year	➢ Determine whether and to what degree students have learned the material taught ➢ Determine annual learning progress and achievement ➢ Evaluate students' annual progress and achievement across groups and jurisdictions ➢ Evaluate effectiveness of educational programs ➢ Measure progress toward improvement goals ➢ Make course-placement decisions • Focus on long-term learning goals and objectives/all material covered in a course, program, or school year

➢ = Common purpose/use

Overview of Relevant Assessment Models

Assessment	Administration	General Purposes and Uses
Formative	A wide variety of methods that teachers use to conduct in-process evaluations of student comprehension, learning needs, and academic progress during a lesson, unit, or course	• Identify concepts that students are struggling to understand and/or skills they are having difficulty acquiring • Focus on short-term learning goals and objectives/material recently covered • Collect detailed information that can be used to improve instruction and student learning *while it is happening*; information is collected and used to adjust teaching on an ongoing basis
Interim	Administered at different intervals during a school year Typically fall between formative and summative assessments Separate from the process of instructing students	• Determine whether and to what degree students have learned material taught • Focus on short- and intermediate-term learning goals and objectives/material covered over a short period of time • Evaluate student progress and achievement across groups, typically at the classroom, building, and/or district levels ◦track trends ◦track student performance in relation to likelihood of performance on other future assessments

Major Design Elements



2

2 Major Design Elements

Section Learning Objectives

Understand that through-year assessments involve multiple inferences about students for multiple purposes.

Characterize key distinctions in how the **content domain** can be "structured" across assessment "modules".

Examine practical issues related to **administration**.

Articulate an overall framing for **score aggregation** as well as provide detail on a **variety of approaches**.

Context

- Through-year assessment programs **adds an additional aim or purpose, or purposes,** into the state summative assessment system.
  - Likely, this means that states are taking on an additional role, or roles, in student learning **previously left to local educational agencies**.
- Designing and implementing a high-quality assessment that serves a single purpose is challenging.
  - Designing and implementing through-year assessments that fulfill multiple purposes has, and will continue to, be **very challenging**.

Considering Through-Year Design

- Through-year assessment programs require much greater:
  - **tailoring to context** and
  - **investment of resources**

  than typical summative assessment programs.
- There is no one "through-year assessment"—there are **many possible designs**.
  - Each user's context and purposes will guide which design is more suitable.
- No one has yet produced a design or program that has been **widely accepted** as meeting educational, political, technical, and feasibility constraints for general statewide assessments.

Interpretive and Validity Arguments

Through-year assessment programs are meant to **support multiple inferences about student performance** for **multiple purposes and uses**, and thus will require multiple interpretive and validity arguments[1] (Kane, 2006; Bennett, Kane & Bridgeman, 2011).

These arguments will need to support:

- A **summative inference** meant support annual determinations under ESSA, and
- One or more **additional inferences** that support the additional purposes
  - These additional inferences may be quite numerous and varied within a particular through-year program, as well as between various through-year programs

[1]This could be in the form of multiple interpretive and validity arguments, or a single argument with multiple parts. The distinction is not important, but articulating the inference and supporting evidence is.

Through-year assessment programs are meant to support multiple inferences about students in service of multiple intended purposes and uses. **Logic and evidence must be developed and gathered to support these inferences and uses**.

	Theory of Action	Interpretive & Validity Argument
Intended Design	• Inputs • Actions • Outcomes	• Inference • Design Elements
Supporting Evidence	• Are the actions implemented with fidelity? • Do they lead to the intended outcomes?	• Are the inferences about what students know and can do supported?

Framing Drawn from Bennett, Kane & Bridgeman, 2011

Interpretive & Validity Argument

- Inference
- Design Elements

Summative Inference	**Logic of the Interpretive Argument**	**Evidence Summarized in a Validity Argument**
Additional Inference #1	**Logic of the Interpretive Argument**	**Evidence Summarized in a Validity Argument**
• • •		
Additional Inference #N	**Logic of the Interpretive Argument**	**Evidence Summarized in a Validity Argument**

Interpretive & Validity Argument

- Inference
- Design Elements

- There are a number of decision points, or **design elements**, that:
  - Distinguish between through-year assessment programs
  - Determine, in part, how well any through-year program will support any given inference and associated purpose and use
- We will explore three of these key design elements:

Content Structure

Administration

Aggregation Method

Design Elements

Content Structure	Administration	Aggregation Method
• How the **content domain** is **organized or structured** across the assessments, which helps define • The **number** and **timing** of the assessments as well as, • The **grain-size** at which the content is allocated to each assessment.	• Whether the assessments are administered in **windows** or **on demand**, as well as • Whether the **order** of the assessments is fixed or flexible, • Which assessments are **required**, and finally, • **Who decides** which assessments are administered and when.	• Whether the single summative score is based on **both within-year and end-of-year results**, or only **end-of-year results**, which is informed by, • **Values** and the **summative claim**, and • Supported by a **measurement model** and **an aggregation method.**

Content Structure

- One of the most **important features** is how the content domain is organized across the multiple assessments (or modules), as it impacts almost every other design decision.
  - Following the language within Race To the Top, we use the term “module” instead of assessment to indicate the discrete administrations of assessed content.
- One major distinction is whether the full domain[1] is represented in each module, or whether a subset of the domain is represented.
  - There are several **unique considerations** when a subset of the domain is used, related to administration, comparability, alignment and score aggregation.

[1]Prior authors have referred to this distinction as “same vs. different blueprint” (Dadey & Gong, 2017) or “distributed” (Gianopulos, 2019).

Content Structure

- Decision Points
- Considerations

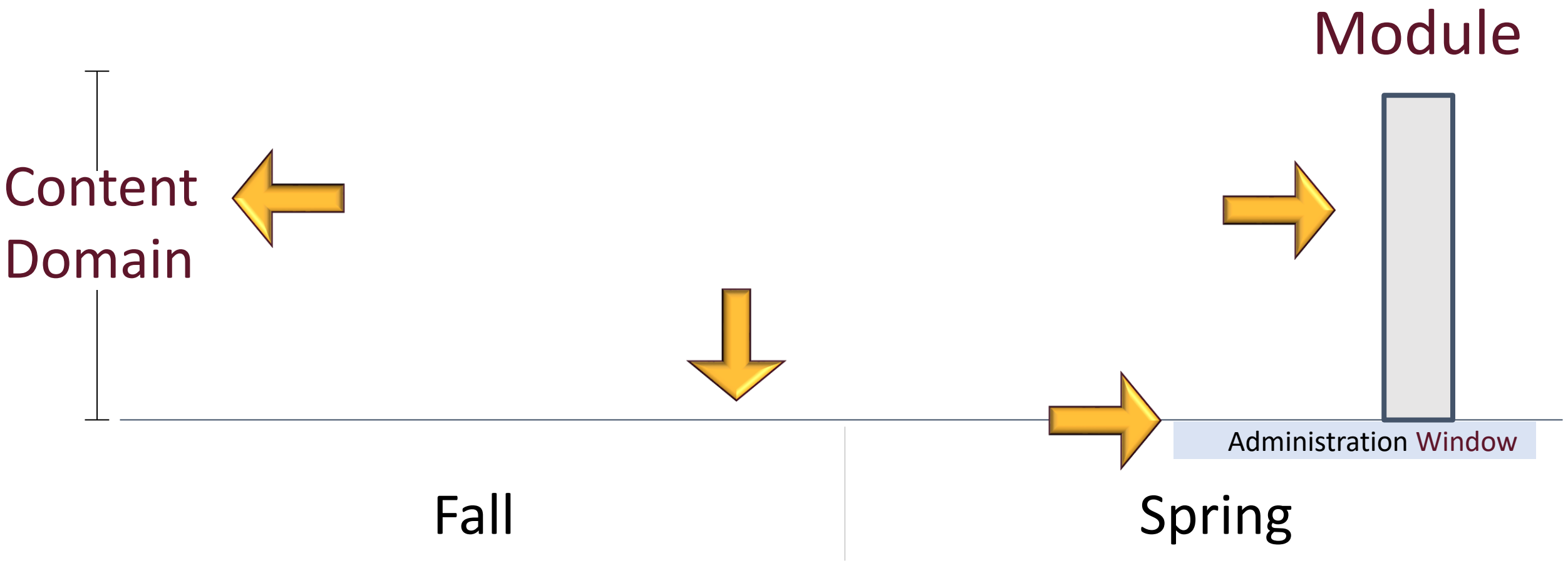
- How the content domain is defined, divided, and articulated across the assessment "modules" is based on:
  - the **aspect or "structure"** of the content domain that is used to **allocate content to each module** (e.g., are the divisions based on the standards, progression or curriculum?) and
  - the **grain-size** at which content is allocated to each assessment module,
  - both of which help define the **number** and **timing** of the assessment modules, as well as,
  - the **flexibility**, or lack thereof, in the ordering and timing of the assessment modules[1]

[1]We also address flexibility in the administration section.

A Hopefully Helpful Heuristic

Typical End of Year Summative Testing

Content Structure

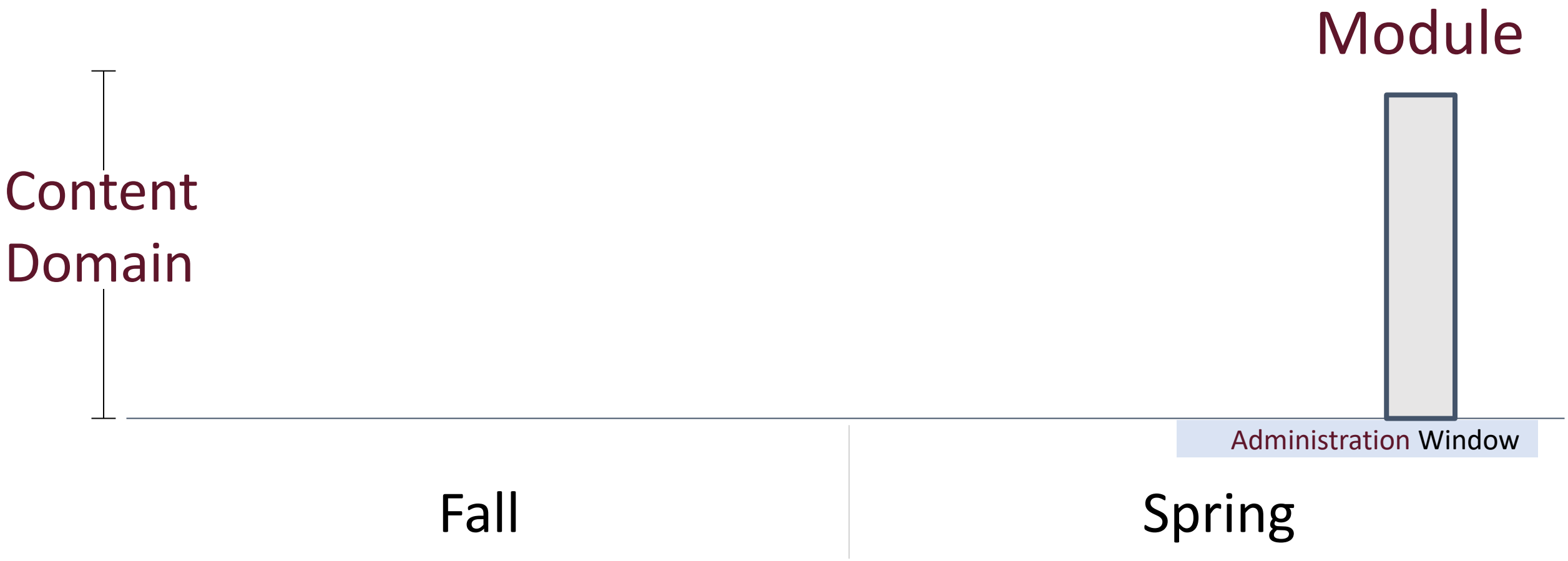


From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

A Hopefully Helpful Heuristic

Typical End of Year Summative Testing

Content Structure

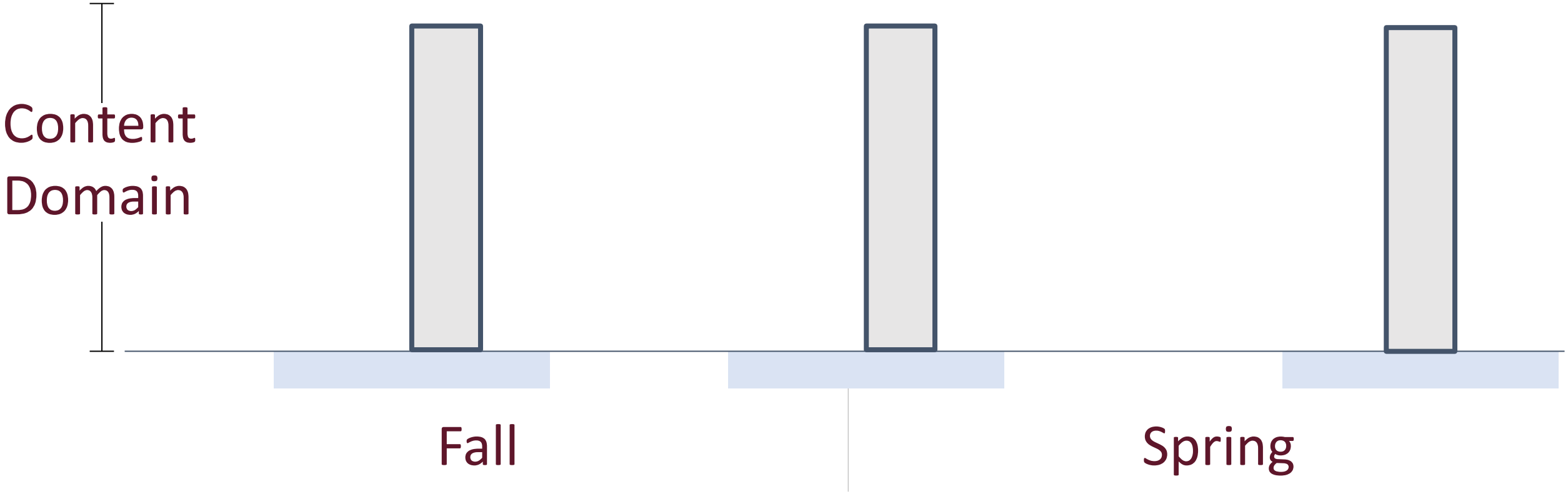


From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

“Modular Full-Domain” Design

Content Structure

Each module (a) covers the entire content domain and (b) is identical in terms of content coverage

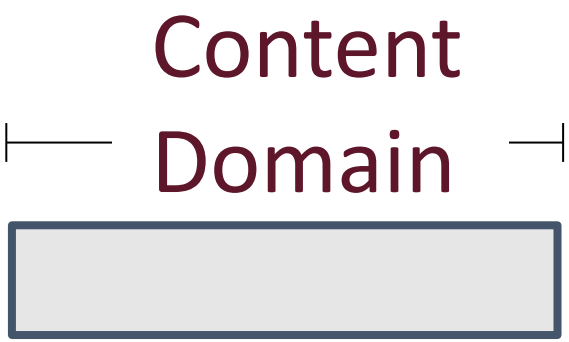


From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

An alternative design is to distribute – or **modularize** – the content domain across modules.

The design question then becomes how to do so.

Content Structure

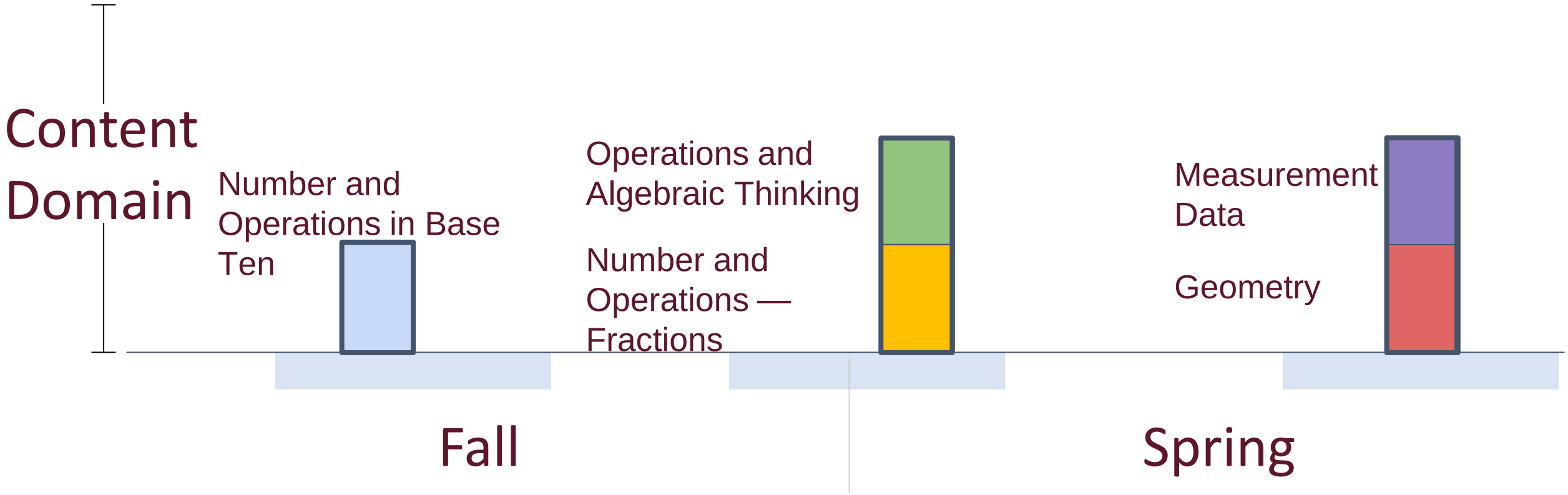


From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

“Modular Sub-Domain” Design

Content Structure

Each module covers a unique group of standards defined by a topic domain

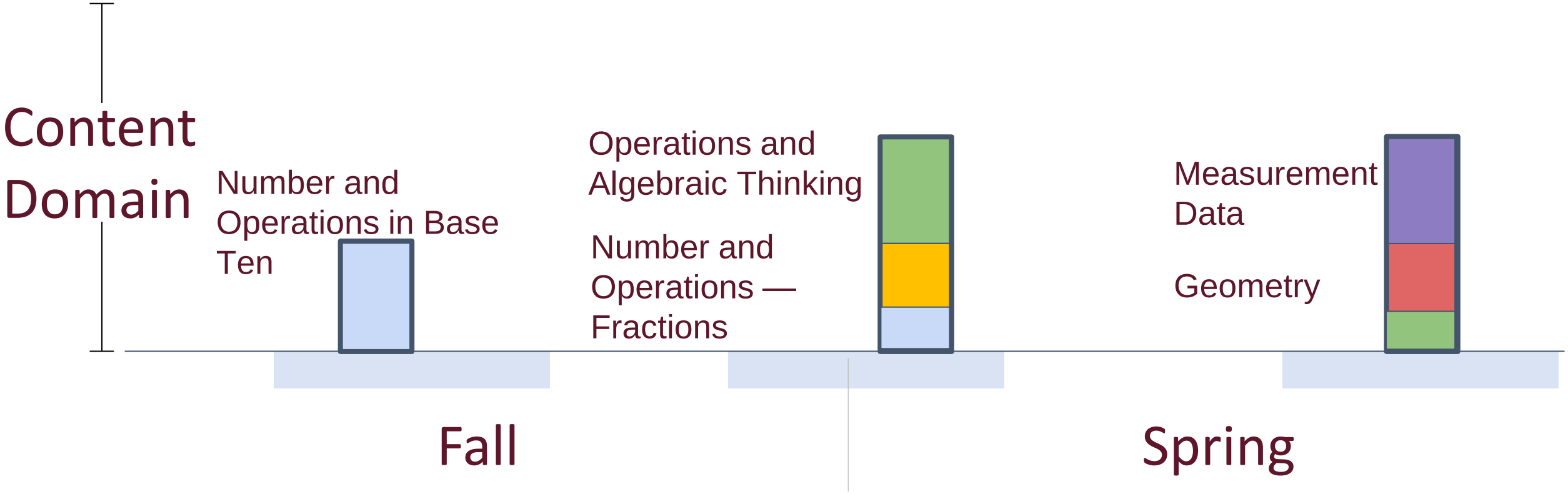


From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

“Modular Overlapping Sub-Domain” Design

Content Structure

Each module covers a unique group of standards defined by a topic domain



From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

Content Structure

“Modular Overlapping Sub-Domain” Design

Each module covers a unique group of standards defined by a topic domain

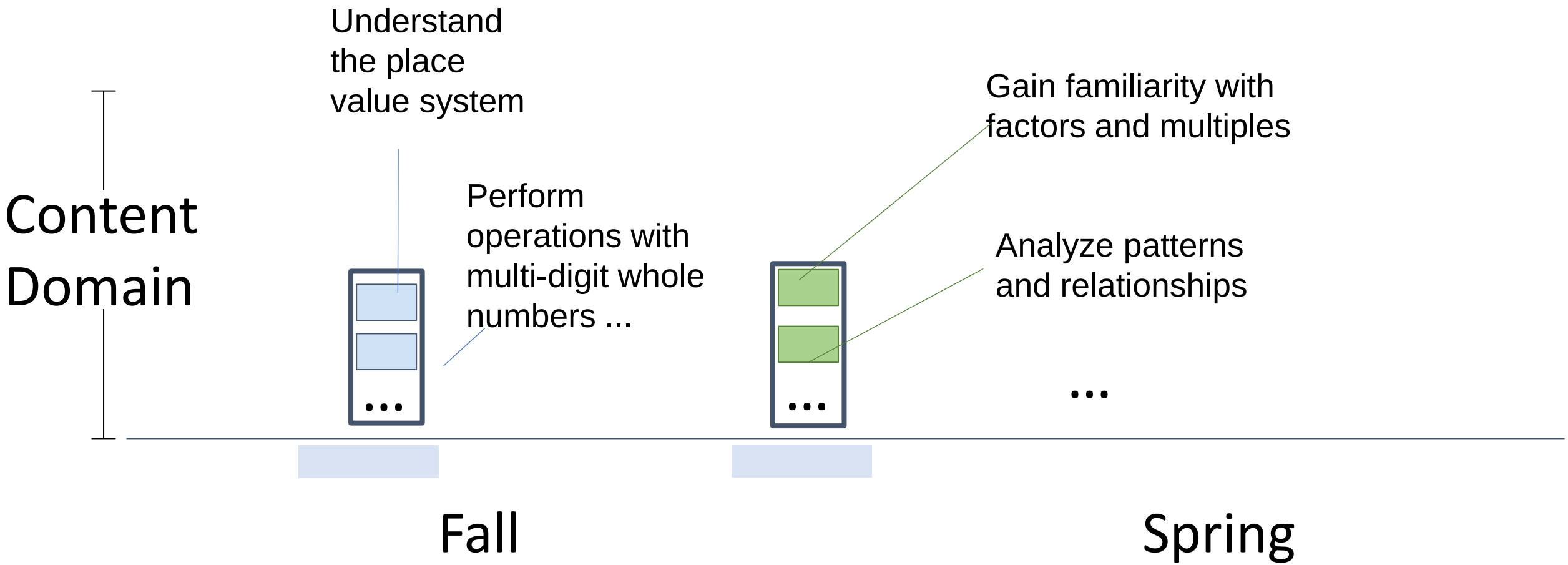
When each module captures a **subset of the domain**, the question is not only how content is allocated, but **how the modules are administered during the year**, and whether there is flexibility in administration. This includes:

- Is the order of administration fixed, or flexible?
  - If it is flexible, who decides (e.g., teacher, leader, state)?
- Can modules be grouped into a single administration?
  - If so how?

“Modular Standards” Content Approach with Grouped Administrations

Content Structure

Each module covers an individual standard and are administered in groups

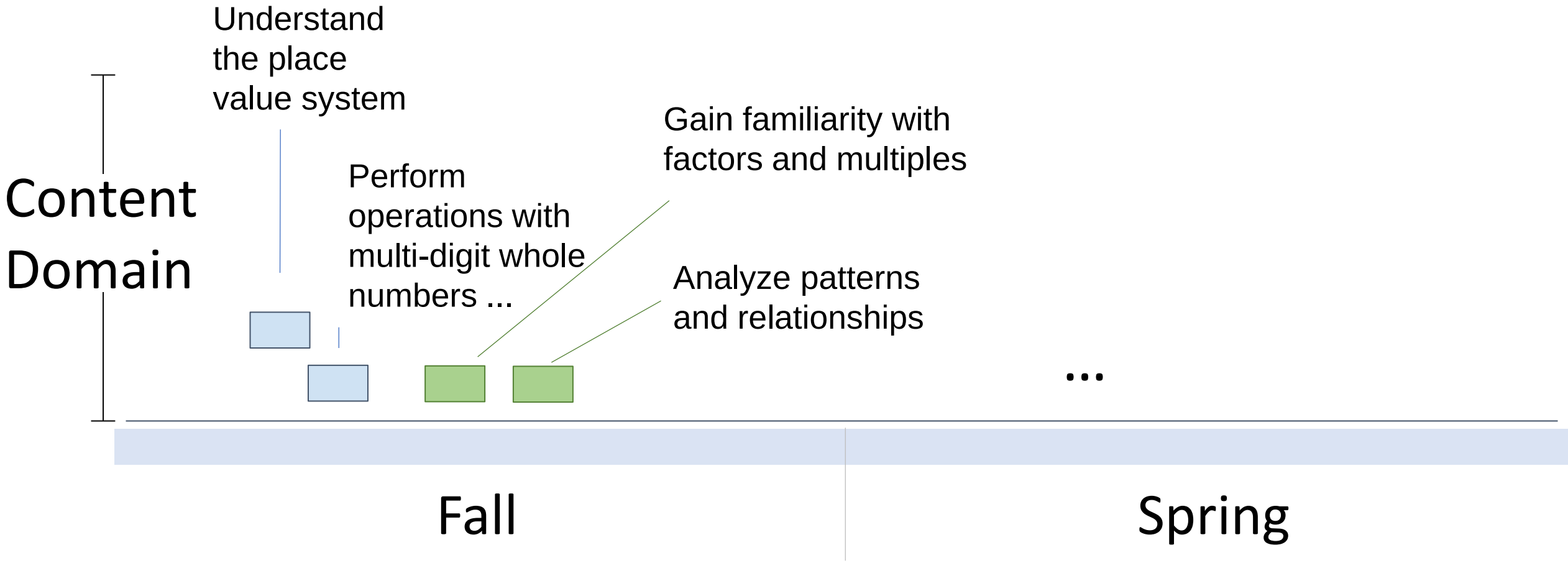


From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

“Modular Standards” Content Approach with Flexible Administrations

Content Structure

Each module covers an individual standard and are administered flexibly based on teacher judgment



From Dadey, Gong, Lorié, & Marion (2021) and Dadey & Gong (2023). Used with permission.

Content Structure

- Decision Points
- Considerations

- Many designs have used the **standards** to allocate content to each module.
  - But learning does not occur in **discrete chunks that align to standards**.
- Alternate approaches to allocating content include:
  - Curricular scopes and sequences (e.g., sequences of modules, on per curriculum)
  - Research based learning progressions, when available
  - Models of complexity or sophistication

Design Elements

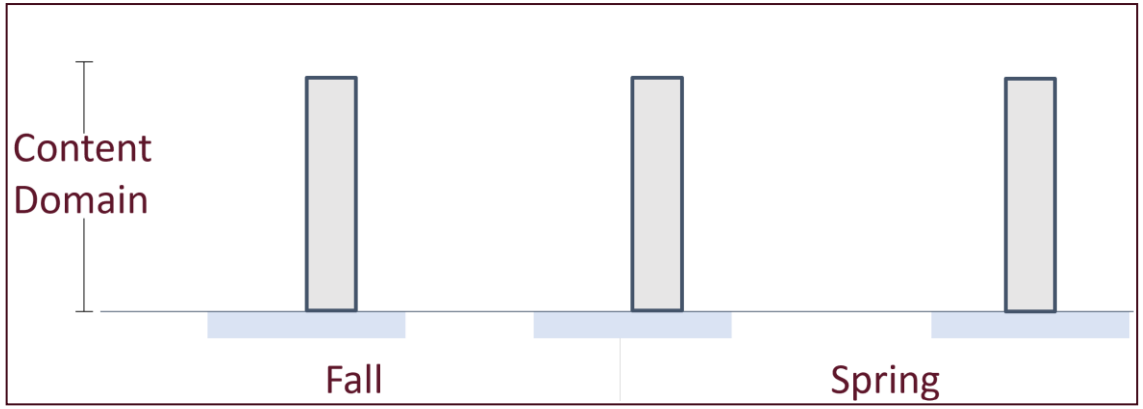
Content Structure	Administration	Aggregation Method
• How the **content domain** is **organized or structured** across the assessments, which helps define • The **number** and **timing** of the assessments as well as, • The **grain-size** at which the content is allocated to each assessment.	• Whether the assessments are administered in **windows** or **on demand**, as well as • Whether the **order** of the assessments is fixed or flexible, • Which assessments are **required**, and finally, • **Who decides** which assessments are administered and when.	• Whether the single summative score is based on **both within-year and end-of-year results**, or only **end-of-year results**, which is informed by, • **Values** and the **summative claim**, and • Supported by a **measurement model** and **an aggregation method.**

Administration

- Structure
- Logistics
- Missing Data
- Accommodations

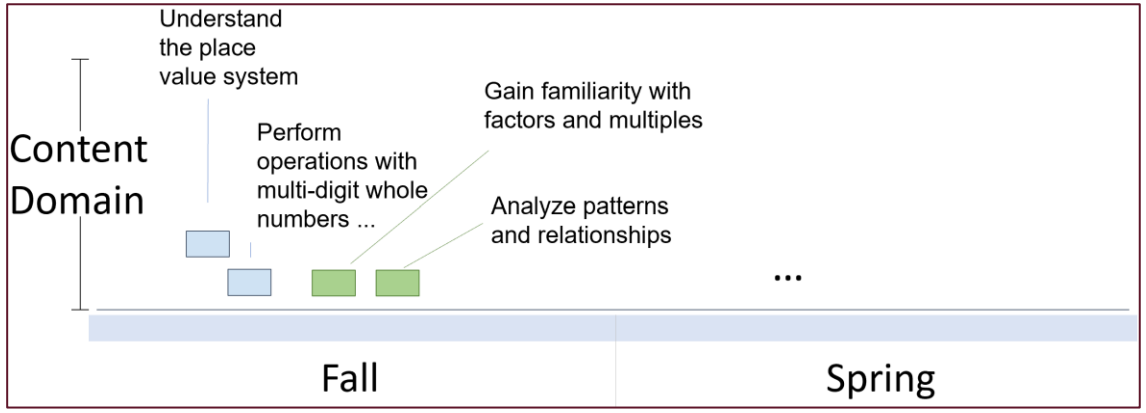
- The content structure, in part, **determines** how the assessments will be administered.
  - The flexibility and complexity of the content structure corresponds to the number of options that can be considered for administration.

Full Domain



Limited choices involving what modules are given, and when.

Modular Designs

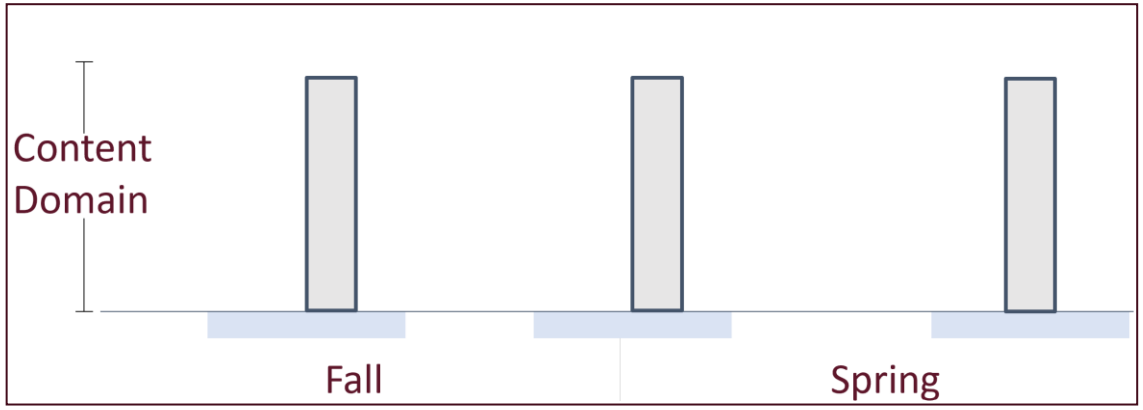


Many more choices involving what modules are given, and when.

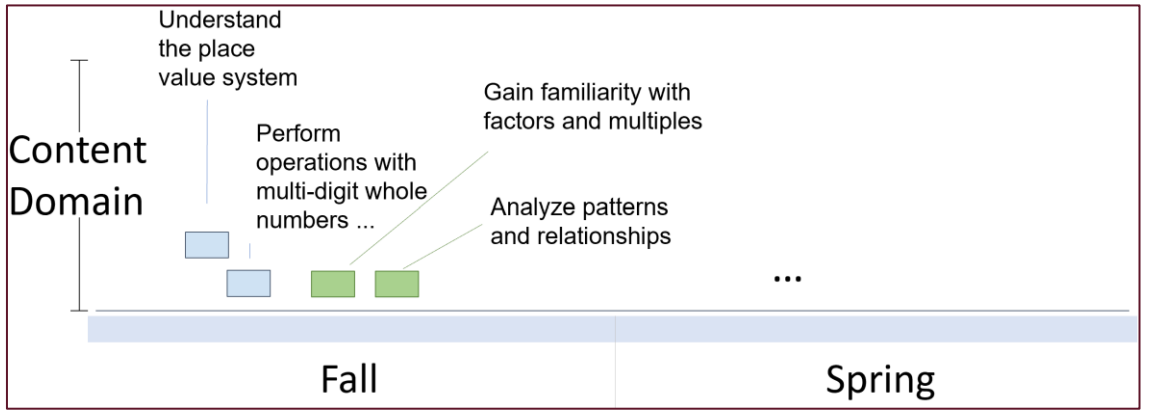
Administration

- Structure
- Logistics
- Missing Data
- Accommodations

Full Domain



Modular Standards



- Whether the assessments are administered in **windows** or **on demand**
- Whether the **order** of the assessments is **fixed or flexible**
- Which assessments are **required**
- **Who decides** which assessments are administered and when (i.e., governance)

Administration

- Structure
- Logistics
- Missing Data
- Accommodations

Increased burden on administrators and coordinators

- **Managing administrations** (e.g., in a state of "always testing")
  - Potentially involving new responsibilities for teachers, leaders, administrators and coordinators
  - Often, district level training on assessment begins in early summer, requiring the appropriate lead time from the program.

- **Understanding and supporting these systems**, particularly if the through-year model is **different across subject areas** (both for those who support administration and those who have to understand it, e.g., principals, district test coordinators)

Administration

- Structure
- Logistics
- Missing Data
- Accommodations

- Addressing missing data involves a combination of
  - **Administration Policy** designed to obtain as complete and administration as possible, including whether make ups are permitted, and if so:
    - When make ups are administered (during or at the end of year)
    - Who makes decisions around make ups
  - **Rules or adjustments** that addressing missing data, including rules around what constitutes a valid score for accountability (i.e., what defines participation and non-participation)
    - Whether there is an allowable number of missed administrations
      - And associated rules with missing administrations (e.g., reduced reporting)
    - Special rules for mobile students (e.g., students who enter the state testing system mid-year, students who change grades mid-year, students who move from one school or district to another)

Administration

- Structure
- Logistics
- Missing Data
- Accommodations

- Accommodations must be provided for each administration window, requiring
  - Training
  - Support
  - Monitoring
- Accommodations are part of a larger set of concerns involving how well through-year works for any given student group. Some key questions include:
  - Does the timing and structure of administration work equally well?
  - Are the intended goals reached equally well?

Design Elements

Content Structure	Administration	Aggregation Method
• How the **content domain** is **organized or structured** across the assessments, which helps define • The **number** and **timing** of the assessments as well as, • The **grain-size** at which the content is allocated to each assessment.	• Whether the assessments are administered in **windows** or **on demand**, as well as • Whether the **order** of the assessments is fixed or flexible, • Which assessments are **required**, and finally, • **Who decides** which assessments are administered and when.	• Whether the single summative score is based on **both within-year and end-of-year results**, or only **end-of-year results**, which is informed by, • **Values** and the **summative claim**, and • Supported by a **measurement model** and **an aggregation method.**

Aggregation Method

- Annual determinations are based on “single summative scores” (i.e., scale score, classification or both).
- The creation of a **single summative scores** involves:
  - not only the application of an **aggregation method**[1],
  - but also consideration of **values** and corresponding purposes.
- Aggregation is as much an exercise **in determining what is valued** as it is an **exercise in measurement**.
  - Aggregation also interacts with the way in which the content domain is structured across the modules.

[1]Here we include both the application of a **measurement model** as well as additional post hoc steps like taking the maximum score.

Aggregation Method

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?

Inferences

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?

Aggregation

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.

Value Judgements & Inferences: End of Year

- One way to understand this complexity is through comparisons to classroom grading. E.g., is a class grade based on:
  - The sum of unit tests?
  - On a single end of year test?

Implicit in many considerations of the creation of a single summative score **is that performance at the end of the year** should serve as the basis of the claim. However, end of year performance is one of amongst a number of options.

Aggregation Method

This view is not based in federal requirements, but rather on an understanding that doing so aligns the time of assessment to grade-level standards.

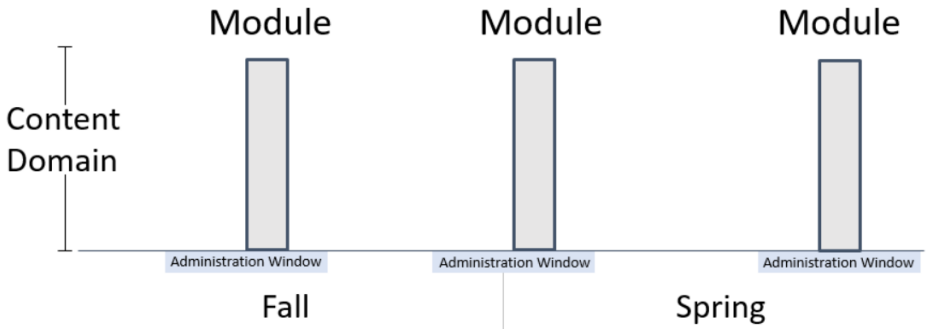
Value Judgements & Inferences: Across the Year

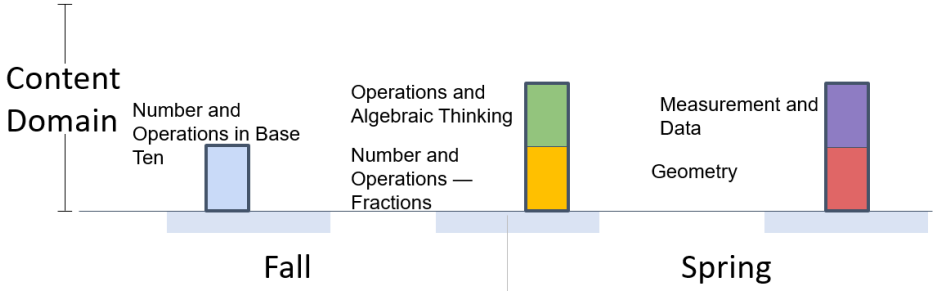
- Alternatively, the inference could be restructure so that it is *not* rooted in the end of the year, e.g., an inference about:
  - "Typical" student performance across the year?
  - "Best" student performance across the year?
- Doing so means explicitly considering the value attached to each module and its associated portion of the content domain.
- Critically, states are currently required to create summative scores that represent the "full set" of grade-level standards.
  - If a set of standards is only assessed once, results from the associated module very likely must be included in the creation of the single summative score.

Aggregation Method

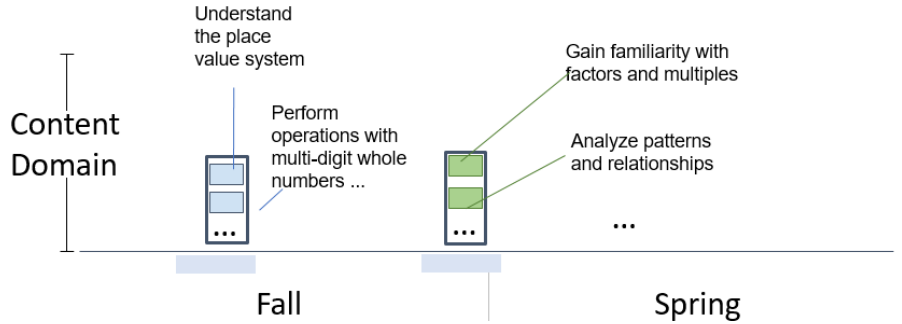
Supporting End of Year Inferences

Supporting end of the year claims based on designs that only measure a chunk of content once within year is difficult.





Increasing Complexity In Supporting End of Year Claims (due to the interaction of content and time)



Aggregation Method

Aggregation Methodology

Measurement Model & Score Creation

- Item Response Theory Models
  - "Traditional Models" calibrated on end of year or through year data
  - Complex models (e.g., multidimensional models, conditioning models)
- Cognitive Diagnostic Models

- Estimation of a latent trait or profile

AND

- "Simple" Aggregation Rules
  - sum, average, weighted average, maximum
- "Complex" Aggregation Rules
  - Rules akin to those use to produce accountability indices (e.g., status and within year-growth; conjunctive rules)

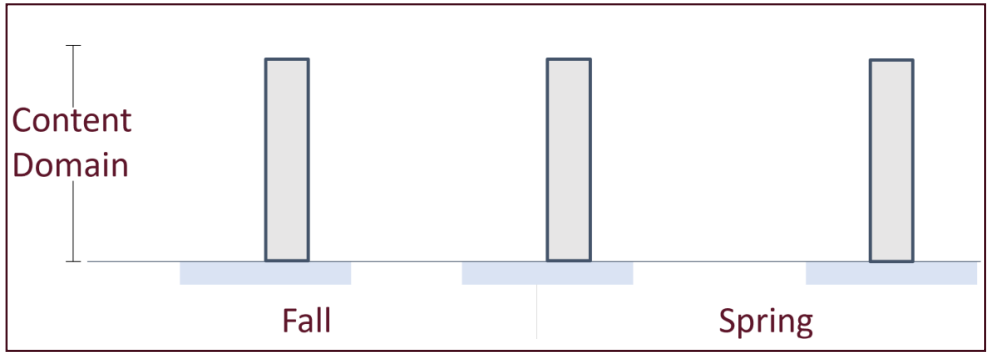
The State of the Field: IRT Based Models

Use a previously calibrated IRT model to:

Preserve End of Year Claims

- Route students within a final module
- Condition student estimates based on the final module using previous score
- Provide subscores based on within year information

Full Domain



Support Across Year claims

- "Simple" operations on scores from each module (sum, average, weighted average, maximum)
- "Complex" operations (e.g., composites that look like accountability indices)

Aggregation Methodology

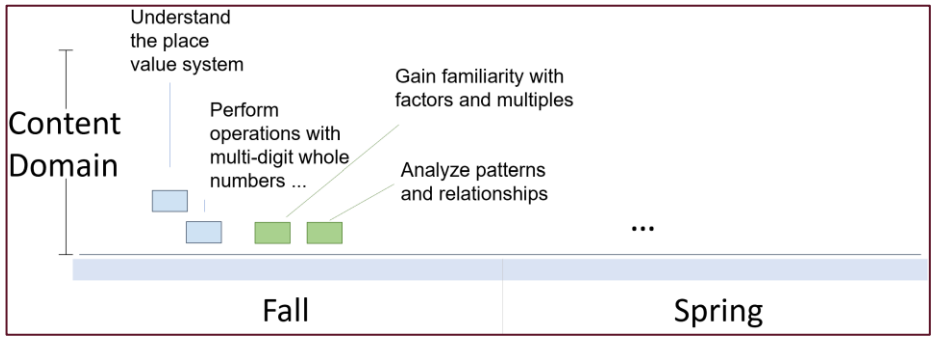
The State of the Field: CDM Based Models

Estimate and use a CDM to:

Preserve End of Year Claims

- Likely cannot support end of year claims, as typically implemented

Modular Standards



Support Across Year claims

- "Simple" operations on scores from each module (sum, average, weighted average, maximum)
- "Complex" operations (e.g., composites that look like accountability indices)

Aggregation Methodology

Aggregation Methodology

Table of Approaches

Summative Score	Aggregation Approach	Example Inference
IRT Based Scale Score	Last Module with precision improved via across module adaptive routing or using prior modules as priors in score estimation	Student proficiency at the end of the year
Weighted Average of IRT Base Scale Scores	An average or weighted average of the scale scores from multiple modules across the year	Typical student proficiency at multiple points during the year
IRT Based Scale Score based on Pooled Data	Treat all modules as if they were all one single assessment, scale using IRT to produce a “composite” or average theta and thus scale score	Typical student proficiency at multiple points during the year
Best of end of year module scale score or weighted average	The single summative score is the best of either (1) the IRT based scale score from the last module or (2) the weighted average across modules	Best of end of the year proficiency or typical student proficiency
Sum of mastered attributes from CDM	Count the number of mastered attributes from CDM based mastery profiles based on standards or skill aligned modules administered on demand	Number of mastered skills, potentially immediately after instruction

Note: The table presents approaches that are currently in use or planned. Other approaches could be possible – this table is not comprehensive.

Examples, Key Challenges, and Methods to Address the Challenges

3

3 Examples, Key Challenges, and Methods to Address the Challenges

Section Learning Objectives

- Become familiar with examples of through-year assessments currently in use or in development
- Discuss the examples in terms of major design elements and considerations
- List key challenges to the development and use of through-year assessments
- Discuss methods to address the challenges in support of the assessment's validity

Emerging through-year assessment models

- **Grading model**
- **Adaptive for end-of-year model**
- **Mastery sequence model**
- **Supplemented summative subscore model**

** Note: The naming of the models are being newly proposed by the authors.*
** Caution: Not all models fit into one of the model types.*

Grading model

- Design Elements
- Example

- Simplest and perhaps most familiar model—similar to how a teacher might create a final grade for a class

Model	Content Structure	Administration	Aggregation Method	Selected example
Grading Model	Modular sub-domain and/or full domain; curriculum-integrated	Multiple modules (e.g., at the end of each quarter)	All modules combined into a weighted average	Louisiana*

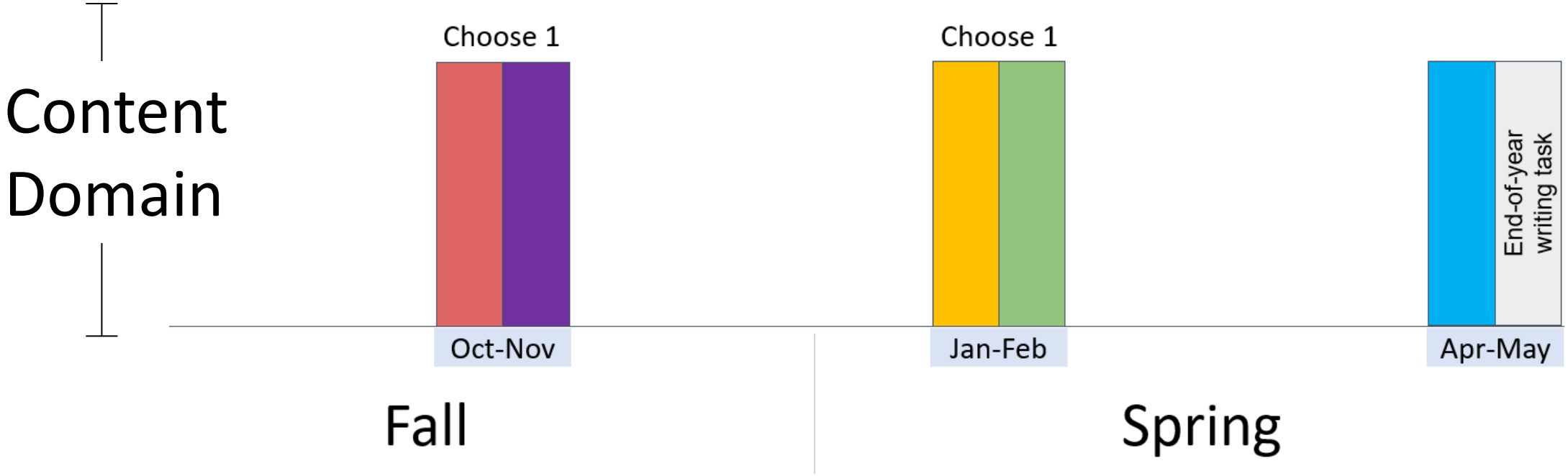
**As of Summer 2023.*

Grading model

- Design Elements
- Example

Louisiana innovative assessment

- The assessment is explicitly aligned with the **curriculum** content.
- All modules are treated as coming from a **single** assessment, essentially treating all modules as equally important, but subject to differential length and discrimination (e.g., using IRT model).



** The figure is an abstract representation of the assessment design and may not correspond with its current operation.*

Adaptive for end-of-year model

- Design Elements
- Example

- This model embodies a decision to downweight the importance of within-year modules. Within-year modules are used to inform the starting point of an adaptive end-of-year module.

Model	Content Structure	Administration	Aggregation Method	Selected example
Adaptive for End-of-Year Model	Modular sub-domain for within-year modules and full domain for an end-of-year module	Multiple within-year modules (optional) and one adaptive end-of-year module (required)	Only end-of-year module is used	North Carolina*

**As of Summer 2023.*

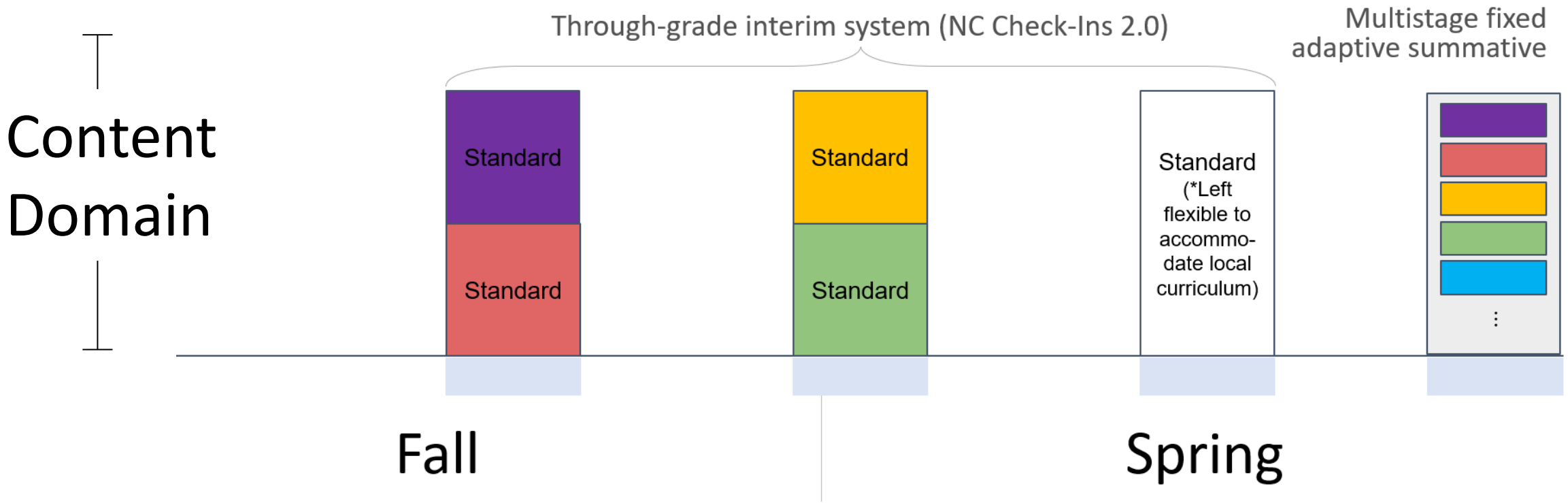
- Maximizes **measurement precision and efficiency** along the entire scale of performance
- Addresses the challenge of **missed attempts** by using information in an **accumulative** manner

Adaptive for end-of-year model

- Design Elements
- Example

North Carolina Personalized Assessment Tool

- Within-year modules and an end-of-year module are ***loosely connected;*** Even when students miss one or more of the within-year modules, they *can* and *must* take the end-of-year module.



** The figure is an abstract representation of the assessment design and may not correspond with its current operation.*

Mastery sequence model

- Design Elements
- Example

- An assessment is designed around a **learning progression model**: It is assumed that a student must have knowledge/skill of certain level *before* acquiring the next level of knowledge/skill.

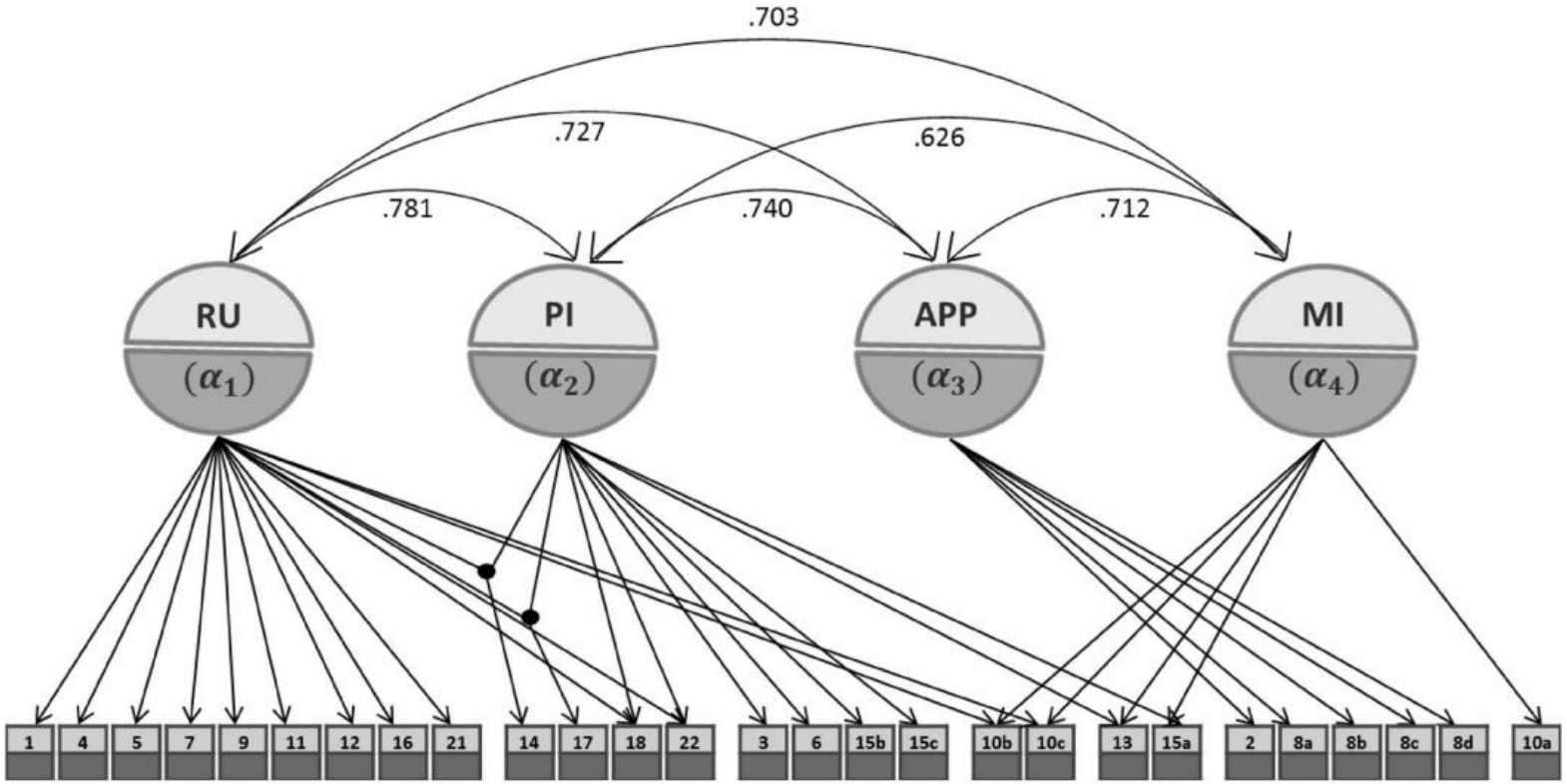
Model	Content Structure	Administration	Aggregation Method	Selected example
Mastery Sequence Model	Each unit of modules is specifically targeted at a knowledge/skill of certain level	Multiple modules of personalized contents, levels, and timing	Profile of the most advanced level achieved in each knowledge/skill component is compiled	Dynamic Learning Maps*

**As of Summer 2023.*

Mastery sequence model

- Design Elements
- Example

- Often, a form of **diagnostic classification model (DCM)** is used to translate item responses into judgments about students' ***profile of mastered skills***—the "highest" level mastered in each components.

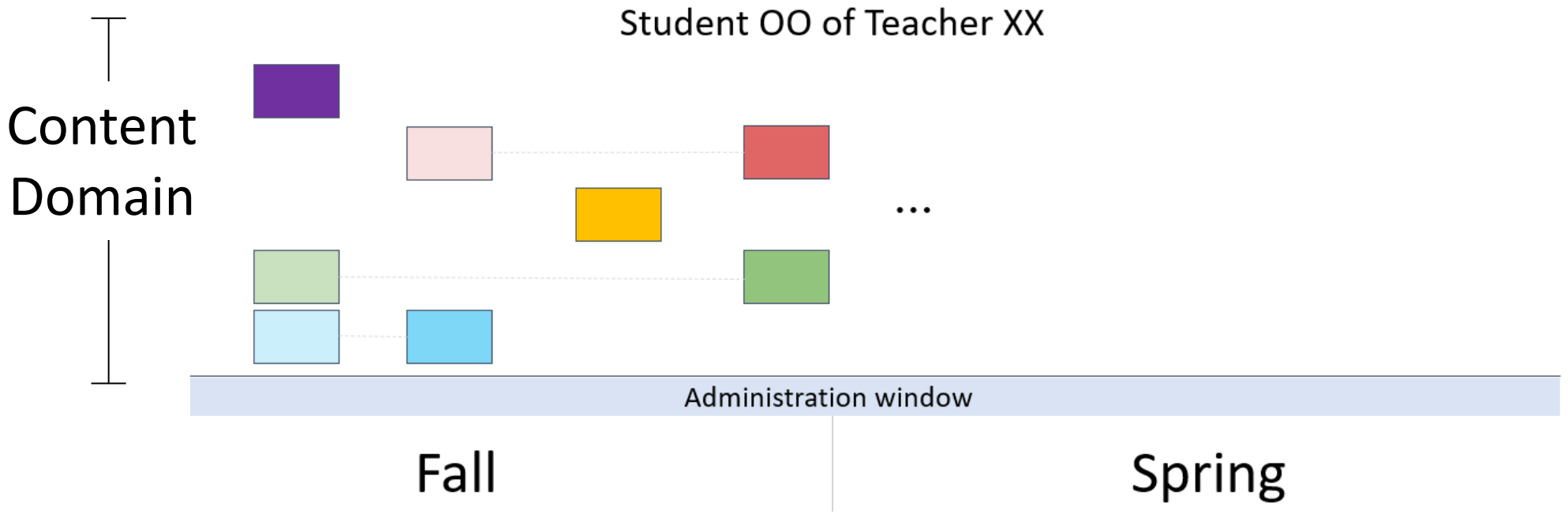


Diagnostic model path diagram (Source: Bradshaw et al., 2014)

Mastery sequence model

- Design Elements
- Example

Dynamic Learning Maps (DLM) Alternate Assessment's instructionally embedded model



** The figure is an abstract representation of the assessment design and may not correspond with its current operation.*

- In each module, teachers must meet the **blueprint requirements**: Given system-provided recommendations, they can freely select the **contents**, **content level**, and **delivery** for each student.
  - Balances between **instruction-assessment alignment** via local flexibility and **content coverage**

Mastery sequence model

- Design Elements
- Example

Dynamic Learning Maps (DLM) Alternate Assessment's instructionally embedded model

- Summative "profile" is based on mastery probabilities from *each* content of *each* content level, with application of a mastery threshold value (e.g., 0.8).
  - The ***profile approach*** partially relieves the burden of score aggregation.
- Interested learners are referred to *Clark and Karvonen (2021)* for theory of action for the assessment system, *Kobrin et al., (2022)* for implementation fidelity, and *DLM Consortium (2022)* for technical details. (Full reference can be found in bibliography.)
- Similar *profile approach* can be found in NAVVY assessment system, one of Georgia's two pilot programs:

  (i) Patterns of standards achieved

  (ii) numerical summary of the standards achieved (e.g., percentage of standards mastered) are combined.

Supplemented summative subscore model

- Design Elements
- Example

- Summative determination is derived from the end-of-year module only; Within-year modules help provide **supplemental subscores**.
  - The supplemental subscores serve formative purpose by providing more fine-grained, contextualized, and instructionally relevant feedback.

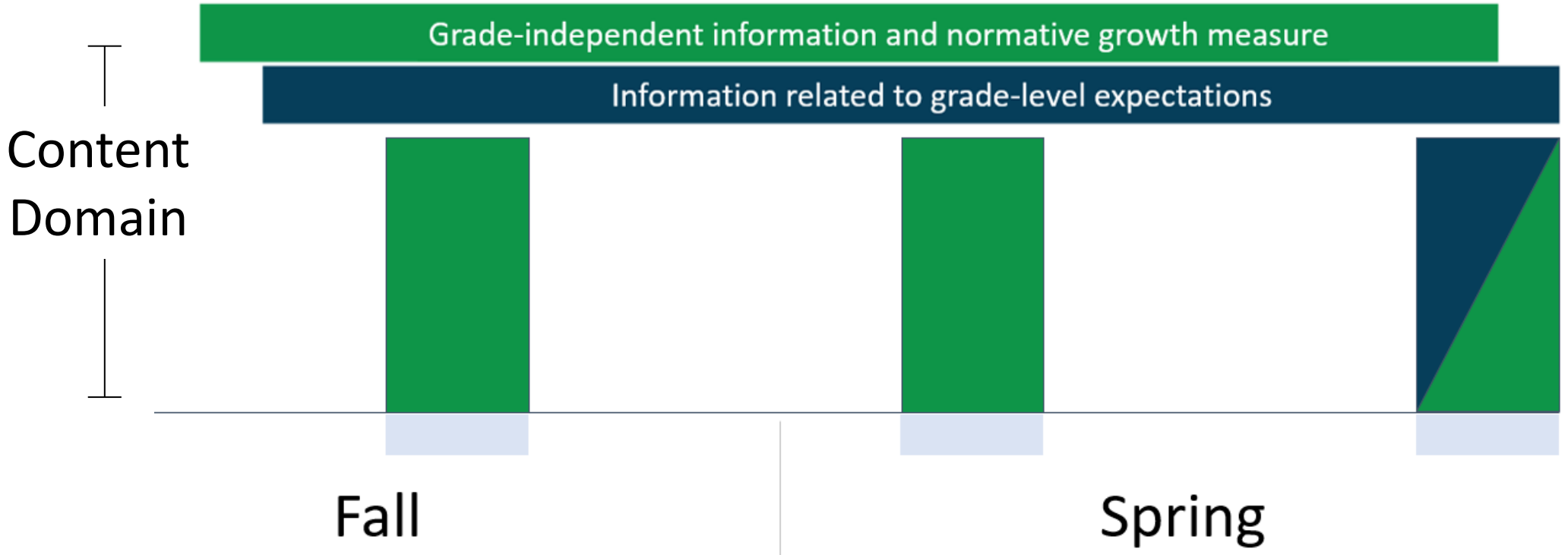
Model	Content Structure	Administration	Aggregation Method	Selected example
Supplemented Summative Subscore Model	Modular sub-domain for within-year modules and full domain for an end-of-year module	Multiple within-year modules for formative purpose; One end-of-year module for summative purpose	Summative score from end-of-year module only; Subscores from both within-year and end-of-year module	Alaska*

**As of Summer 2023.*

Supplemented summative subscore model

- Design Elements
- Example

Alaska innovative assessment system



** The figure is an abstract representation of the assessment design and may not correspond with its current operation.*

- Within-year modules provide ***grade-independent*** norm-referenced indicators of student performance (projected proficiency)
- End-of-year module provides the same information "plus" ***grade-level*** proficiency score for accountability purpose.

More examples*

The information is adopted from Education First (2023), listed in alphabetical order.

1. Alaska's Map Growth of Alaska System of Academic Readiness (AK STAR)
2. Delaware's Social Studies Through-Course Assessment
3. Florida's Assessment of Student Thinking (FAST)
4. Georgia's MAP (GMAP) consortium
5. Georgia's Putnum consortium system
6. Indiana's Smarter Balanced Interims
7. Kansas' Predictive Interim Assessment & Interim Mini Tests
8. Louisiana's Curriculum-connected ELA through-year assessment
9. Maine's Comprehensive Assessment System (MECAS)
10. Montana's ELA Testlets
11. Nebraska's Student-Centered Assessment System (NSCAS) Growth
12. North Carolina's Personalized Assessment Tool (NCPAT)
13. Texas' Through-Year Assessment Pilot (TTAP)
14. Virginia's Growth Assessments

**Through-year assessment models being tested/operated by States.*

Comparative summary of through-year model choices of States (Education First, 2023; Appendix)

FEATURE	LA[3]	DE	GA[1]	MT[4]	NC	TX	GA[2]	NE	AK	FL	ME	VA	KS	IN
Assesses all grade-level standards each time	✓					✓	✓	✓	✓	✓	✓	✓	✓	✓
Assesses a subset of standards each time		✓	✓	✓	✓									
Syncs test with learning or scope and sequence	✓	✓	✓	✓	✓									✓
Curriculum- connected	✓													
Summative score based on final test					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Aggregates tests to create summative score	✓	✓	✓	✓										
Multi-stage or Phase adaptive					✓	✓	✓	✓						
Item-level adaptive							✓	✓	✓	✓	✓	✓		
Provides more regular data to educators and families	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

1: Georgia's Navvy System; **2:** Georgia's GMAP consortium; **3:** Louisiana's curriculum-connected assessments; **4:** Louisiana is piloting IAAS in math

Summary table of emerging through-year model

Model	Content Structure	Aggregation Method	Administration	Selected example
Grading Model	Modular sub-domain and/or full domain; curriculum-integrated	All modules combined into a weighted average	Multiple modules (e.g., at the end of each quarter)	Louisiana
Adaptive for End-of-Year Model	Modular sub-domain for within-year modules and full domain for an end-of-year module	Only end-of-year module is used	Multiple within-year modules (optional) and one adaptive end-of-year module (required)	North Carolina
Mastery Sequence Model	Each unit of modules is specifically targeted at a knowledge/skill of certain level	Profile of the most advanced level achieved in each knowledge/skill component is compiled	Multiple modules of personalized contents, levels, and timing	Dynamic Learning Maps
Supplemented Summative Subscore Model	Modular sub-domain for within-year modules and full domain for an end-of-year module	Summative score from end-of-year module only; Subscore from both within-year and end-of-year module	Multiple within-year modules for formative purpose; One end-of-year module for summative purpose	Alaska

** Note: The naming of the models are being newly proposed by the authors.*
** Caution: Not all models fit into one of the model types.*

Considerations for Implementation

4

Section Learning Objectives

4

Considerations for Implementation

Identify key considerations for successful implementation in addition to technical design

Identify how previous aspects in the module support needed evaluation and continuous improvement

Successful Implementation Requirements

- Good technical design to support valid, reliable, fair, and useful score interpretations

AND ALSO

- Practical implementation of a test life cycle
- Clear sponsorship, ownership, and governance
- Sustainable human, technical, and financial resources
- On-going policy/political support from multiple stakeholders
- Ability to identify strengths and needs, and evolve as necessary

- Often implementation of a through-year assessment will be more complex and resource-intensive than a single end-of-year state summative assessment

Multiple Tests' Life Cycles

- Through-year assessments consist of during-the-year and end-of-year assessments.
  - Each test purpose/design will have a life cycle involving
    - Theory of Action and Validity argument
    - Specification development and documentation
    - Item and test development
    - Test administration
    - Scoring
    - Scaling, equating
    - Standard setting
    - Reporting
    - Quality assurance
    - Analyses to support validation, effective use, etc.
    - *Often technical manual has a chapter for each major topic*
  - Having multiple tests requires having multiple, different test life cycles of development, implementation, revision
  - Need to be coordinated adds complications

Example: Complex Test Life Cycle

- Requires almost **continuous analysis**, particularly in the **first year** if the through-year program is not based on an already existing test program (i.e., scale; some programs are leveraging extant scales)
  - This breaks the "end of year" analysis model in place for most states
  - **Increases the number of FTEs** needed to support the program, both from the vendor and state side
  - Usually increases reliance on vendors
- Piloting and field testing are more extensive, more complicated and more important for program success
  - Particularly if the program is new and lacks a scale

Example: Field Testing in Test Life Cycle

Field Testing

- **Temporal Anchoring.** To which season is a test's scale anchored?
  - One of them
  - All of them
- **FT-OP Season Match.** Do a test's field and operational administration seasons need to match? Will it be OK to administer items in a season other than the one in which they were field tested?
- **Optimizing for Equating.** How should new forms or items be distributed across seasons to support successful equating? (While controlling data collection burden)

Sponsorship, Ownership, Governance

States should clearly identify **sponsorship**, **ownership**, and **governance** for all aspects of the through-year assessment program, several of which may differ from an end-of-year summative program. **States may need to resolve any conflicts.**

Sponsorship, Ownership, Governance

States should clearly identify sponsorship, ownership, and governance for all aspects of the through-year assessment program, several of which may differ from an end-of-year summative program. States may need to resolve any conflicts.

Sponsorship

Who commissions the assessment, determines the intended construct, audience, uses, approves major design decisions, etc.

- The state is typically the sponsor of the state assessment

Sponsorship, Ownership, Governance

States should clearly identify sponsorship, ownership, and governance for all aspects of the through-year assessment program, several of which may differ from an end-of-year summative program. States may need to resolve any conflicts.

Sponsorship

Who commissions the assessment

Ownership

Who owns and controls the use, modification, and branding of items, tests, scales, score names, data, etc.

- Take especial care if considering using an existing commercial assessment

Sponsorship, Ownership, Governance

States should clearly identify sponsorship, ownership, and governance for all aspects of the through-year assessment program, several of which may differ from an end-of-year summative program. States may need to resolve any conflicts.

Sponsorship

Who commissions the assessment

Ownership

Who owns and controls instrument and data

Governance

Who has authority to determine when, to whom, how the assessment is to be administered and by whom and how data are to be used

- The state typically has governance authority over aspects of an end-of-year state assessment
- The district, school, or teacher typically has governance authority over during-the-year assessments, especially those intended to support classroom, school, or district monitoring of teaching and learning.

Policy Considerations

- Policy considerations for design and implementation
- Policy considerations for adoption and implementation
- Policy considerations for evaluation

Policy Considerations

- Policy considerations for design and implementation
  - Relation of local curricula to within-year content specifications
  - State's security requirements for administration of during-the-year assessments
  - Data sharing/privacy considerations
- Policy considerations for adoption and implementation
- Policy considerations for evaluation

Policy Considerations

- Policy considerations for design and implementation
- Policy considerations for adoption and implementation
  - Is the through-year being built based on an existing program, or is it being developed new?
    - In some cases, the dominant interim assessment program has served as the basis of the through-year assessment program. In these cases one question is how much control the SEA has with the program, as well what happens if the SEA would like to shift vendors.
    - If the through-year is meant to replace current local interim assessments, how will the SEA seek policy support and provide support, in terms of governance, funding, and district data processes?
- Policy considerations for evaluation

Policy Considerations

- Policy considerations for design and implementation
- Policy considerations for adoption and implementation
- Policy considerations for evaluation
  - Considering return on investment: How is a through-year system intended to be better than the status quo, and how will it be evaluated?

Evaluating for Improvement

- Evaluating and improving the through-year assessment program will be strengthened by having a theory of action with well-specified **purposes, uses, and logic model** that is specific enough to inform practice
- Implementation of a through-year model intentionally introduces additional use(s) onto the statewide summative assessment
  - Doing so means that the success of the through-year model is much more contextually based than state summative assessment programs

Evaluating for Improvement

- Evaluating and improving the through-year assessment program will be strengthened by having a theory of action specific enough to inform practice
- Implementation of a through-year model intentionally introduces additional use(s) onto the statewide summative assessment
- Because a through-year assessment model is more complex and touches more educational aspects than an end-of-year summative model, the SEA will need to proactively monitor what is working well and what might need improvement, and respond with appropriate policy, technical, and communication adjustments

Summary and Relevant Areas of Research

5

Summary

1. Context

- Desired areas for improvement and innovation of assessment
- Definition of a through-year assessment program

> *"A through-year assessment program consists of **multiple distinct assessments administered across the school year** where information from the multiple assessments is (i) combined to yield a **summative determination** of student performance to support federally required systems of school identification and support, and (ii) used to **support another purpose or purposes** (e.g., logistical, administrative, or monitoring)."*

- Overview of relevant assessment models (summative/formative/interim)

Summary

1. **Context**
2. **Major Design Elements**
   1) *Content structure.* Selection of content allocated to each module, grain-size of each module, frequency and timing of modules, flexibility in administration of modules, etc.
   2) *Administration*. Understanding/support/management of the program, addressing the issue of missed administrations, etc.
   3) *Aggregation method*. Need of a measurement model (such as IRT or CDM) and a score creation model to estimate a latent trait or a profile of multiple latent traits

Summary

1. **Context**
2. **Major Design Elements**
3. **Examples, Key Challenges, and Methods to Address the Challenges**

- Emerging models:
  - Grading model (e.g., Louisiana)
  - Adaptive for end-of-year model (e.g., North Carolina)
  - Mastery sequence model (e.g., Dynamic Learning Maps)
  - Supplemented summative subscore model (e.g., Alaska)
- Different models have different strengths in addressing key challenges (e.g., missed administration, instruction-assessment alignment, content coverage, score aggregation for summative scores).

Summary

1. **Context**
2. **Major Design Elements**
3. **Examples, Key Challenges, and Methods to Address the Challenges**
4. **Considerations for Implementation**
   - Implementation of multiple tests' life cycle (development-implementation-revision);
   - Identification of sponsorship, ownership, and governance;
   - Securing sustainable human, technical, and financial resources;
   - Policy/political support from stakeholders;
   - Evaluation of the assessment system itself to identify strengths and needs to evolve as necessary, etc.

Relevant areas of research

- Establishing and maintaining item bank:
  - Development/validation of theorized learning progression; development of culturally sensitive/relevant items/tasks for diverse population; monitoring and adjustment for item parameter drift; item security
- Measurement model for longitudinal data:
  - Longitudinal/higher-order diagnostic classification model: useful for profile-based standard setting based on binary constructs (mastery/non-mastery)
  - Multilevel/multidimensional item response theory model: useful for performance level outcome in continuous scale
- Score generation model
  - Use of parameter estimates from a measurement model; threshold of "sufficient information" when students missed one or more administrations

Relevant areas of research (cont.)

- Measure of growth/change
  - Inference on growth/change in student/classroom/school-levels using indices such as student growth percentile; Conditioning/contextualizing scores on opportunity to learn
- Utilization of process data collected from computer/online modules
  - e.g., response time, response changes, number of attempts
- Score reporting
  - Validity and reliability of scores reported in student/teacher/classroom/school/district/state level
  - Interpretation and instructional usage of the reports with caution to avoid confirmation bias and transition into actionable information

You have reached the end of this section…

Dadey, N., Gong. B., Kim, Y., & Sato, E. (2024). Through-year Assessment [Digital ITEMS Module 35]. *Educational Measurement: Issues and Practice, 43*(1),97-98. https://doi.org/10.1111/emip.12595