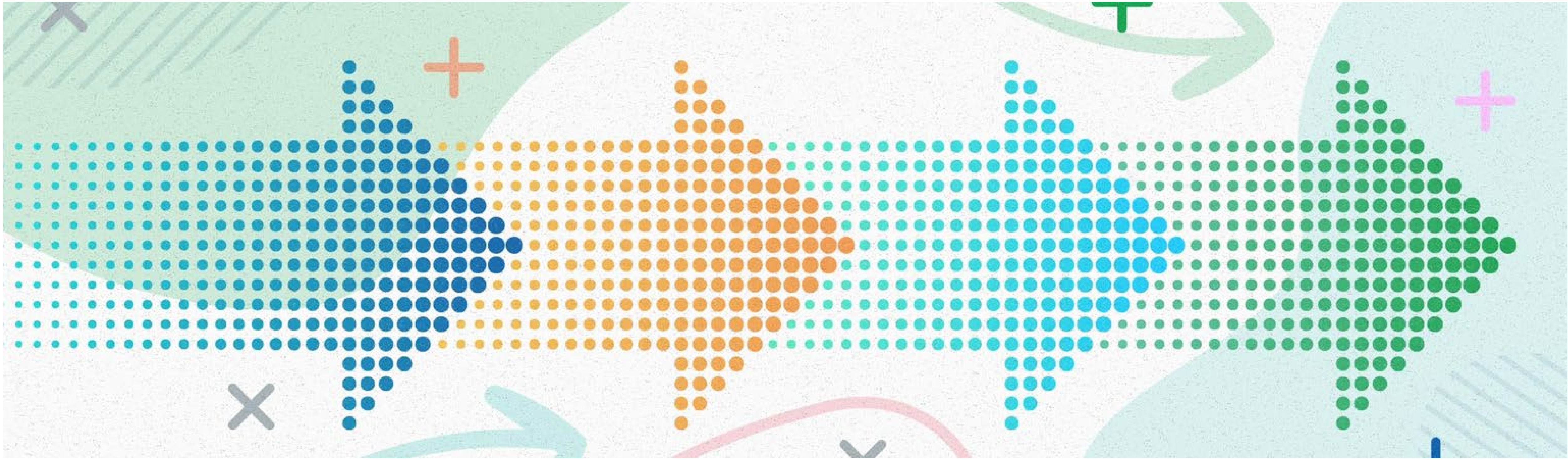


Through-Year Assessment Virtual Convening

November 15-16, 2021

The National Center for the Improvement of Educational Assessment

Center for Assessment

Monday, November 15th, 1-2:30 PM ET

Defining terms, considering aims, and diving into key design features.

Monday, November 15th, 3-5 PM ET

Connecting use cases and claims, and the designs that support them, together to consider needed evidence.

Tuesday, November 16th, 1-2:30 PM ET

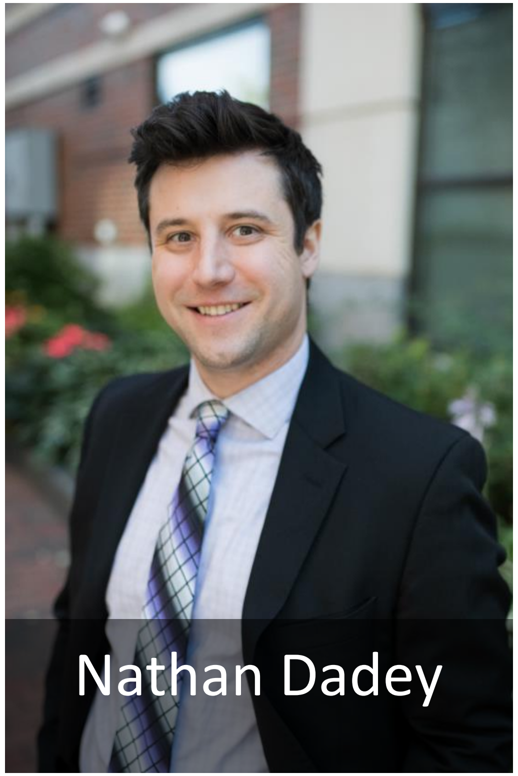
In depth consideration of key big picture technical and logistical issues.

Tuesday, November 16th, 3-5 PM ET

What will it take to make through-year assessment systems work to support students and educators?

The Organizers

Will **Lorié**



Nathan Dadey

Brian Gong

Scott Marion

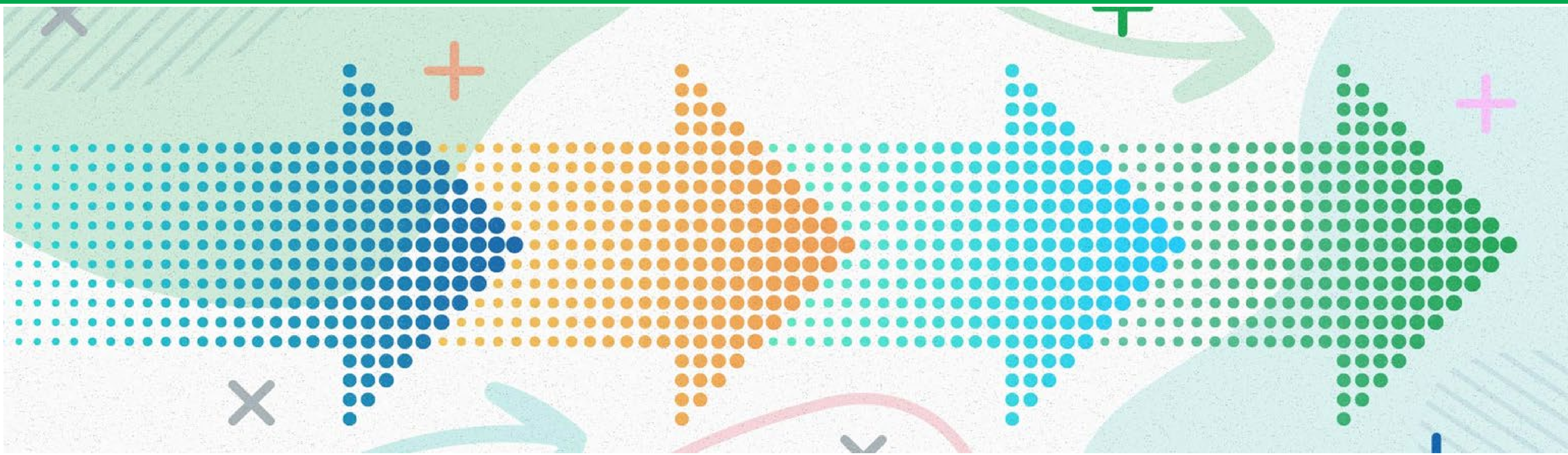
A Brief Recap of Sessions 1 & 2

- **Session 1**—We offered a definition and described a few classes of through-year designs.
- **Session 2**—We took a deep dive into both summative and instructional claims associated through-year designs and described some of the evidence necessary to substantiate such claims.
- **Sessions 3 & 4**—Today!

- All materials will be posted here by the end of the week:
  - https://www.nciea.org/events/claims-and-evidence-through-year-assessments-what-we-know-and-what-we-need-know

Session 3

- We’ve been focusing on definitions, designs, and claims.
- In Session 3 we’re going to deal with some of the thorny questions and issue that keep folks like us awake at night.



Session 3: Technical & Logistical Issues

Through-Year Virtual Convening, November 15, 2021

Will Lorie, Nathan Dadey, Brian Gong, & Scott Marion

The National Center for the Improvement of Educational Assessment

Center for Assessment

Outline

1 **Technical & Logistical Issues**

- Aggregation
- Alignment
- Field Testing
- Standard Setting
- Reporting

2 **Invited Presentations**

In depth considerations from invited participants

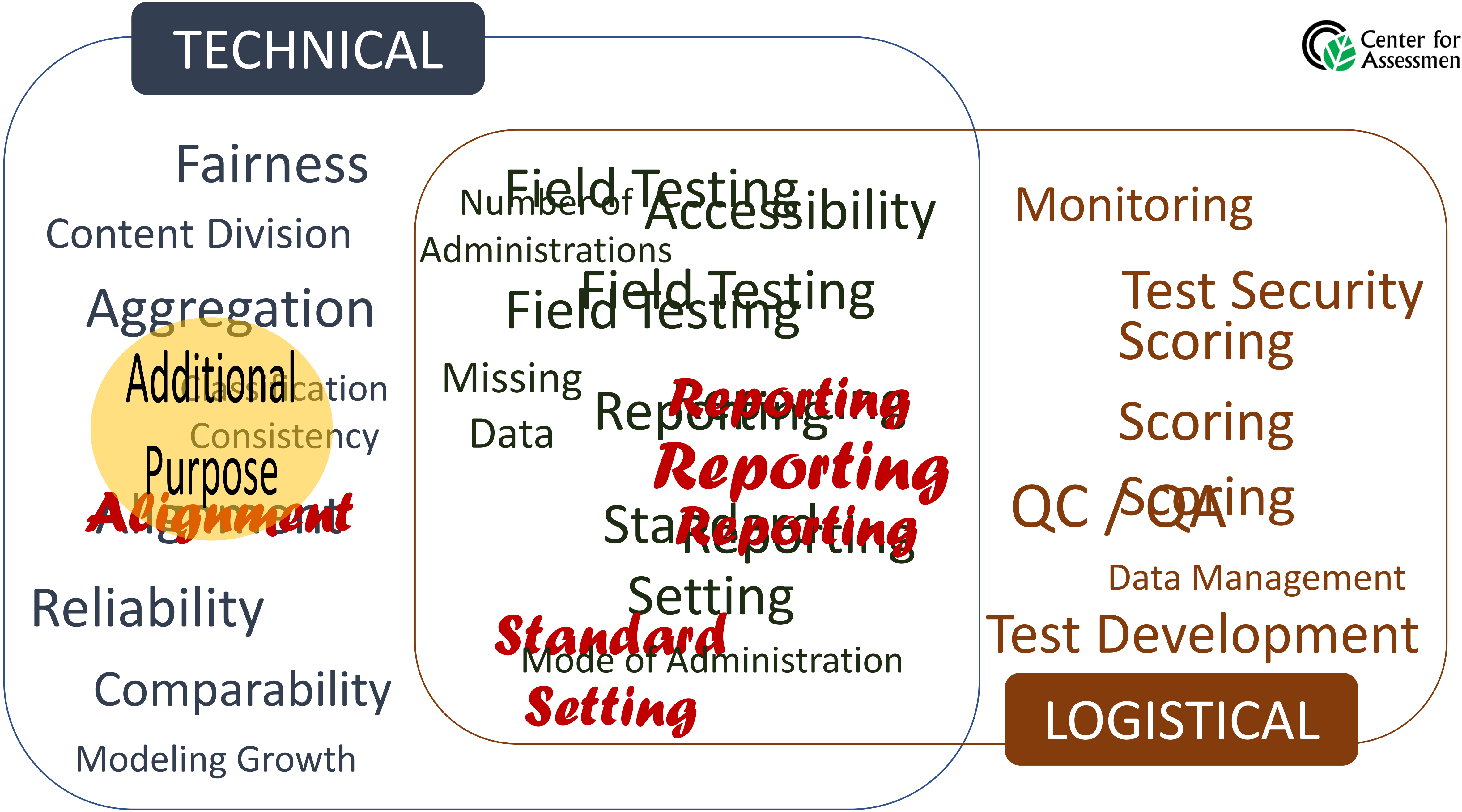
3 **Question and Answer Session**

Facilitated Audience Interaction

1. Technical and Logistical Issues

What technical and logistical issues keep you up at night?

https://pollev.com/cassessment154



TECHNICAL

Fairness

Content Division

Aggregation

Additional Purpose

Classification Consistency

Alignment

Alignment

Reliability

Comparability

Modeling Growth

Number of Administrations

Field Testing

Accessibility

Field Testing

Field Testing

Missing Data

Reporting

Reporting

Reporting

Standard Setting

Reporting

Reporting

Standard Setting

Mode of Administration

Monitoring

Test Security

Scoring

Scoring

QC / QA

Scoring

Data Management

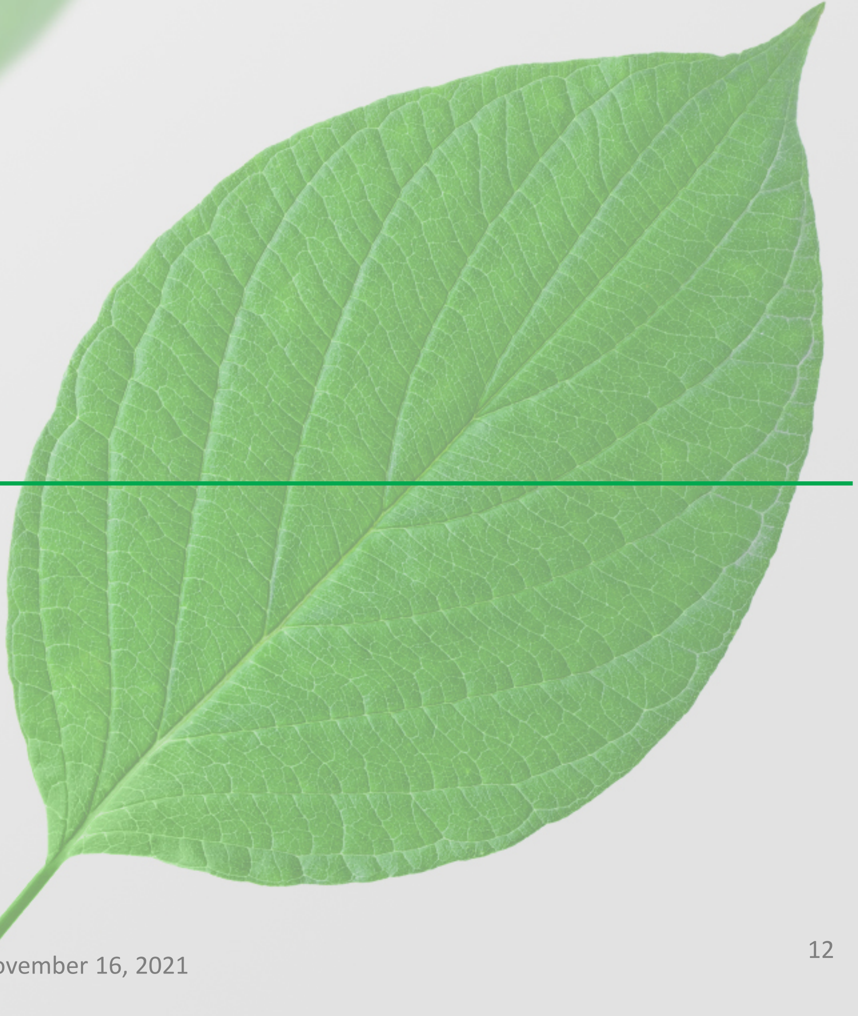
Test Development

LOGISTICAL

Preview

1. Aggregation
2. Alignment
3. Field testing
4. Standard setting
5. Reporting

1. Aggregation



To support annual determinations, we need a **single summative score**.

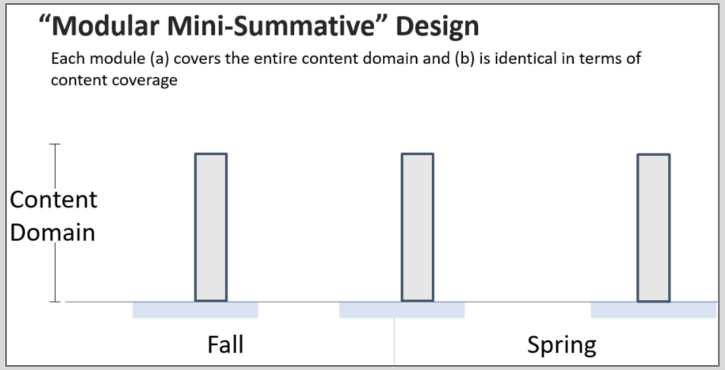
The creation of a single summative scores involves not only the application of an **aggregation method**[1],

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.

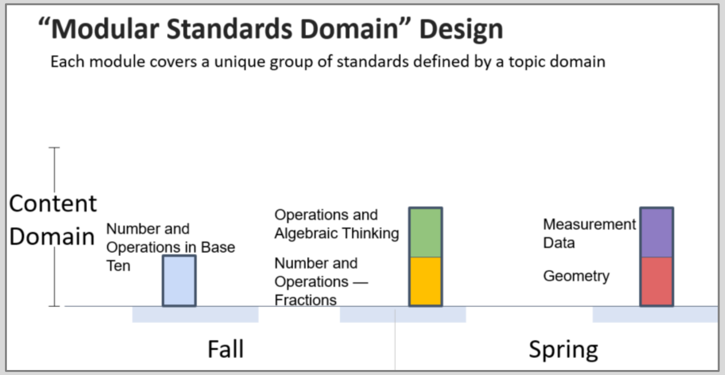
Restated, every aggregation method reflects specific values and supports specific claims.

In Session 2



“Modular Mini-Summative” Design

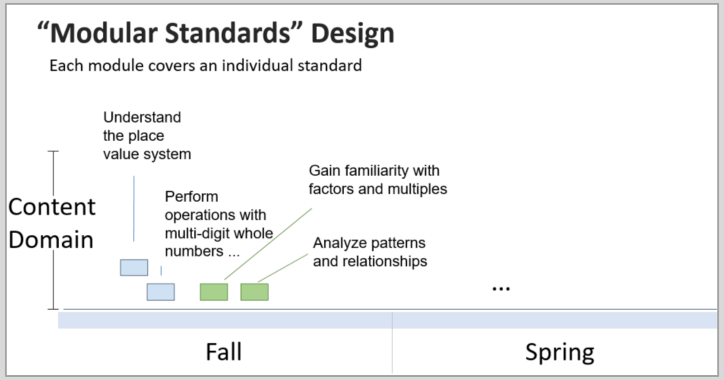
Each module (a) covers the entire content domain and (b) is identical in terms of content coverage



“Modular Standards Domain” Design

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

Content Domain
Number and Operations in Base Ten
Operations and Algebraic Thinking
Number and Operations — Fractions
Measurement Data
Geometry
Fall
Spring



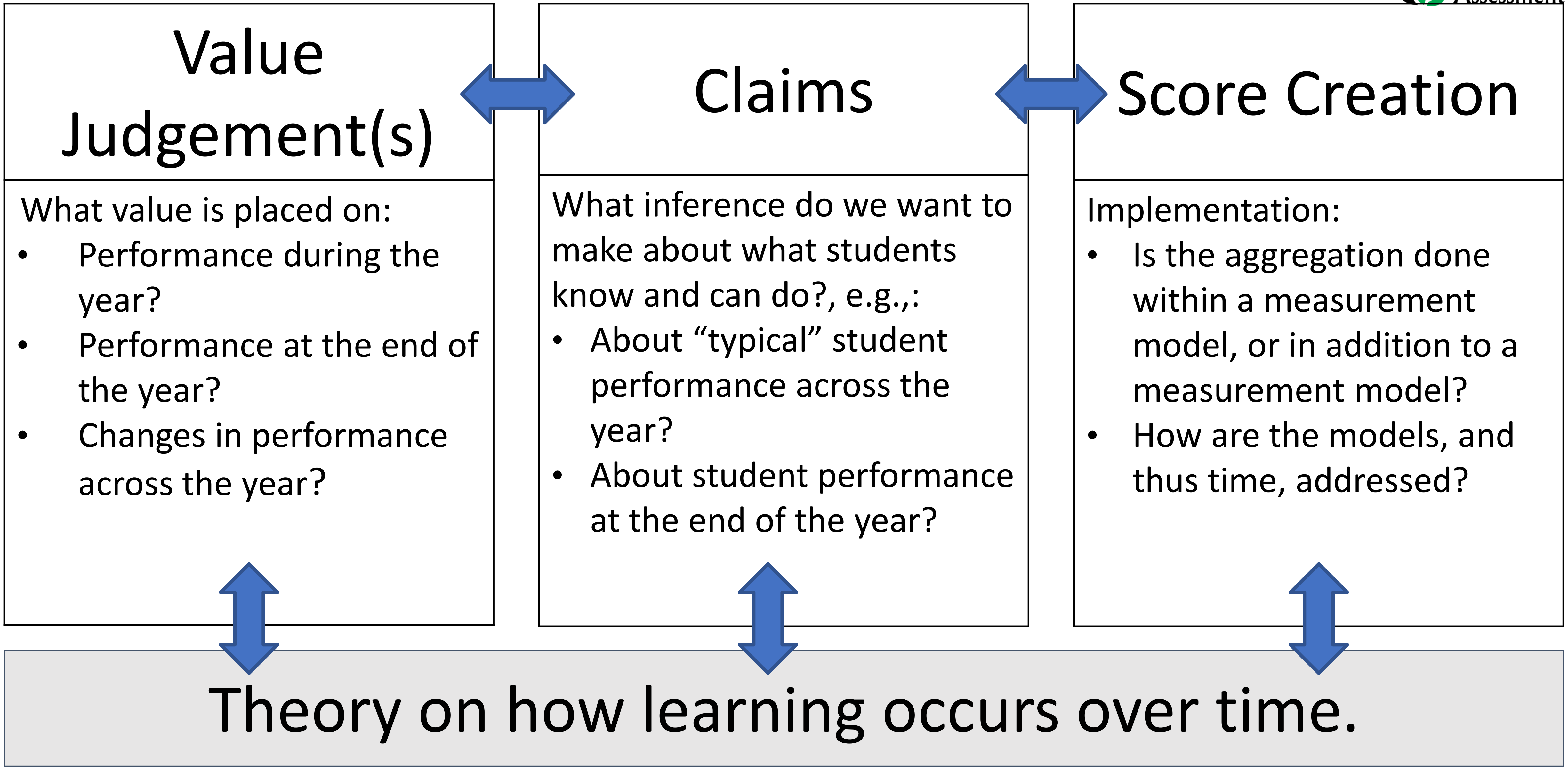
“Modular Standards” Design

Each module covers an individual standard

Content Domain
Understand the place value system
Perform operations with multi-digit whole numbers ...
Gain familiarity with factors and multiples
Analyze patterns and relationships
...
Fall
Spring

Claims in relation to:

- **Content and Administration Design.** Interaction of the distribution of content and administration with time.
- **Intended Inference.** End of the year or “something else”



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.,:

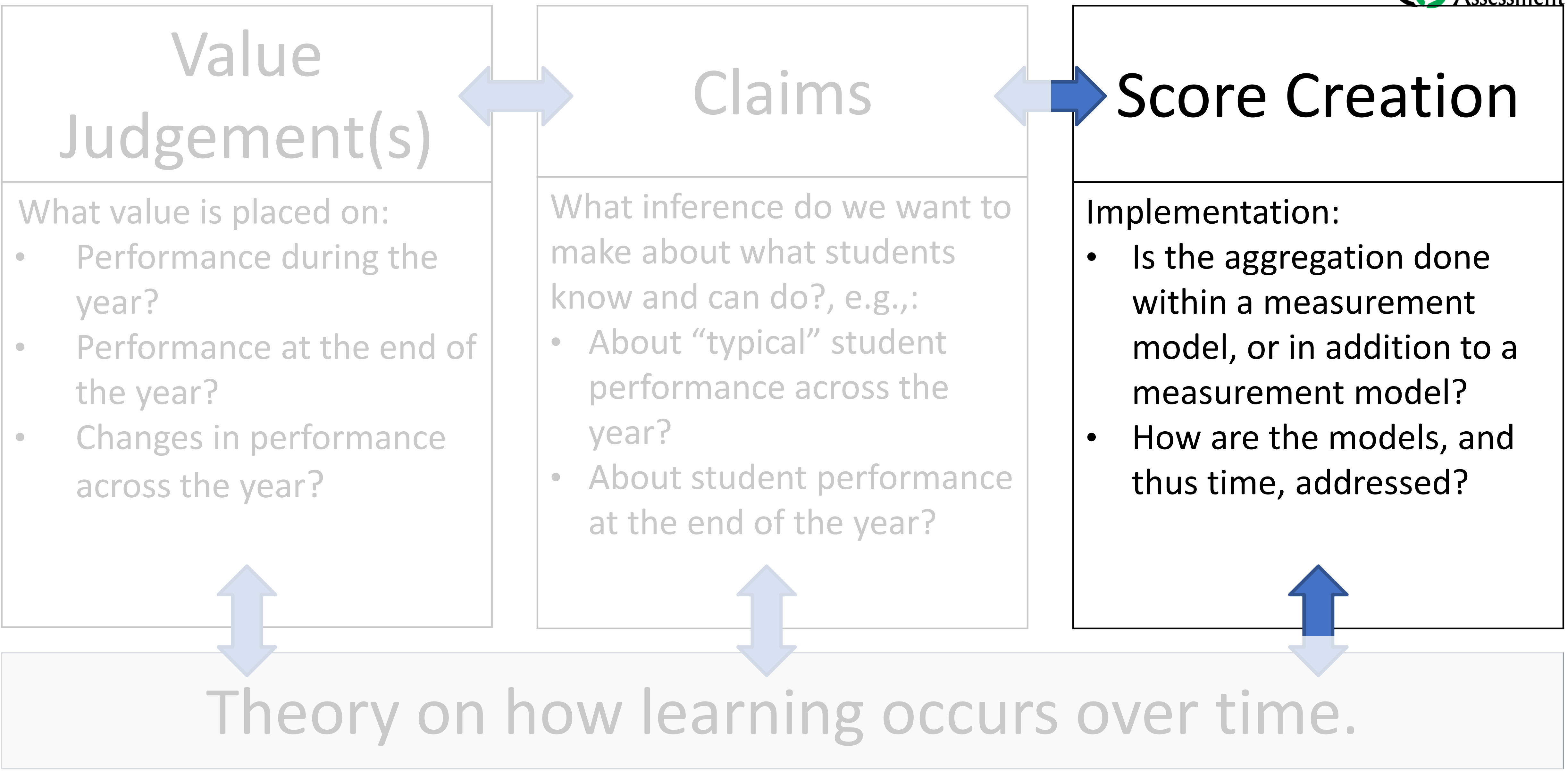
- About “typical” student performance across the year?
- About student performance at the end of the year?

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.



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.

Now let’s dive into the current state of the field by examining three overlapping approaches to measurement models and score creation.

Measurement Model

- 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

See also Gianopulos, 2019 for a summary of proposed models.

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

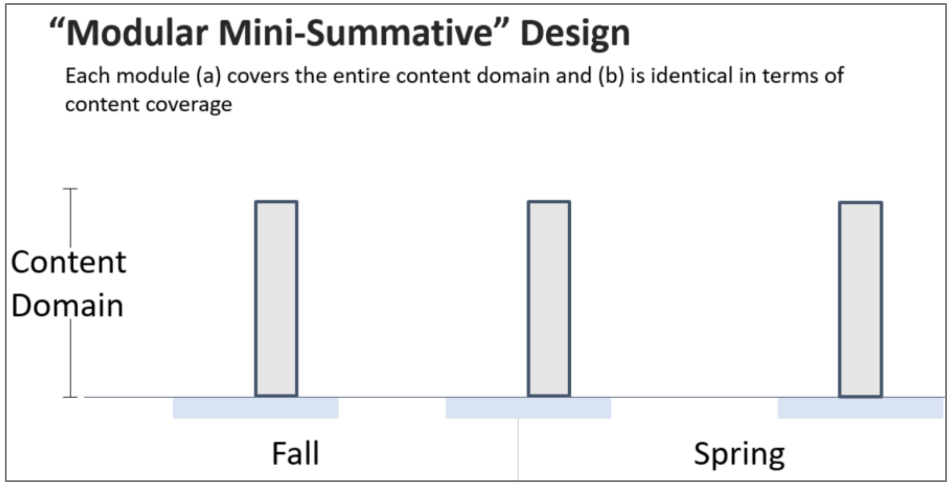
- "Simple" Aggregation Rules
  - sum, average, weighted average, maximum
- "Complex" Aggregation Rules
  - Rules akin to those use to produce accountability indices (e.g., status and within year-growth; conjunctive rules)

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


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

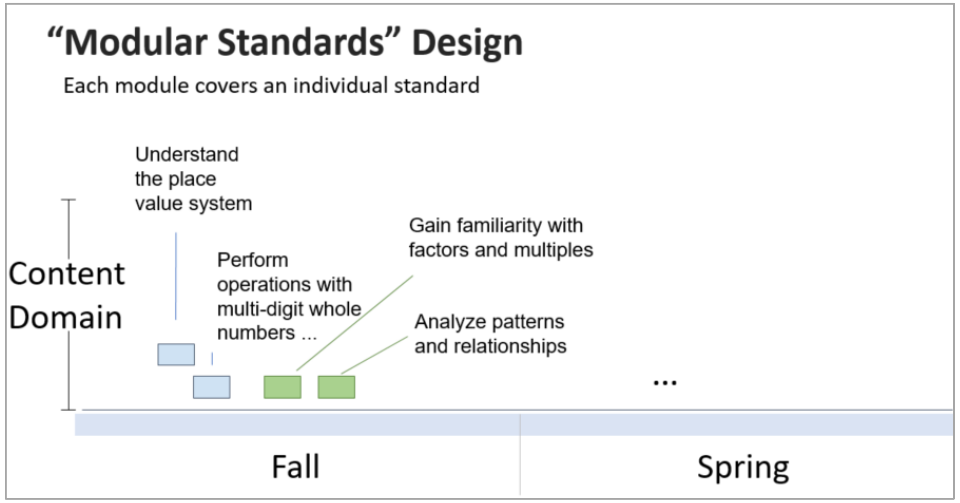
Support Across Year claims

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

The State of the Field: CDM Based Models

Estimate and use a CDM to:

Preserve End of Year Claims



“Modular Standards” Design

Each module covers an individual standard

Understand the place value system
Content Domain
Perform operations with multi-digit whole numbers ...
Gain familiarity with factors and multiples
Analyze patterns and relationships
...
Fall
Spring

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)

Open Questions, PT. I

- How do we understand and **investigate student learning and opportunity to learn (OTL)**?
  - What implications do the patterns of learning and OTL have for the design of the through-year program and subsequent aggregation?
- What is the **range of aggregation methods** and how can we compare them?
- What are the **technical properties of single summative scores**?
  - Measurement precision and error
  - Year to year variability at the aggregate level

Open Questions, PT. II

- How do we decide what should **count** within an aggregation process?
  - E.g., all modules, only specific modules, only parts of specific modules (i.e., items).
- Will different parts of the through-year assessment system be used for **different purposes**?
  - E.g., within a mini-summative model, a single summative score could be based on all three windows, but across year growth calculated just on the last module.
- How can we engage with **stakeholders** about, and explain to, the single summative score?

2. Alignment

Alignment

- The extent to which the test content reflects the depth and breadth of grade level academic content standards
- Largely a technical issue
- Through-year introduces two new challenges:
  - Which test? Each one? The last only? All of them taken together?
  - If the academic content standards represent end-of-year expectations, to what standards should earlier-in-the-year tests align?

Alignment and Through-Year: Open Questions

- **Definition.** What precisely do we mean by *alignment* for non-contemporaneous collections of assessments?
- **Relation to Aggregation.** How do the results of alignment depend on score aggregation methods?

What needs to align to standards – The test(s) a student took or their relative contributions to her score?

- **Evidence.** What constitutes sufficient evidence of alignment? How can through-year program designers incorporate alignment guidance into their planning?

3. Field Testing

Field Testing

Field testing allows us equate test forms, blocks, or items so that assessment programs can be refreshed over time.

- Technical: Equate to when?
- Logistical: When to equate?

Through-Year Field Testing: Open Questions

- **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 successful equating? (While controlling data collection burden.)

4. Standards Setting

Standard Setting

“I start from what's being assessed, intended score interpretations and uses (SIUs), and the PLDs or other definitions of levels of performance. That's the starting place that's common to all principled approaches.” -S. Ferrara

Score 1

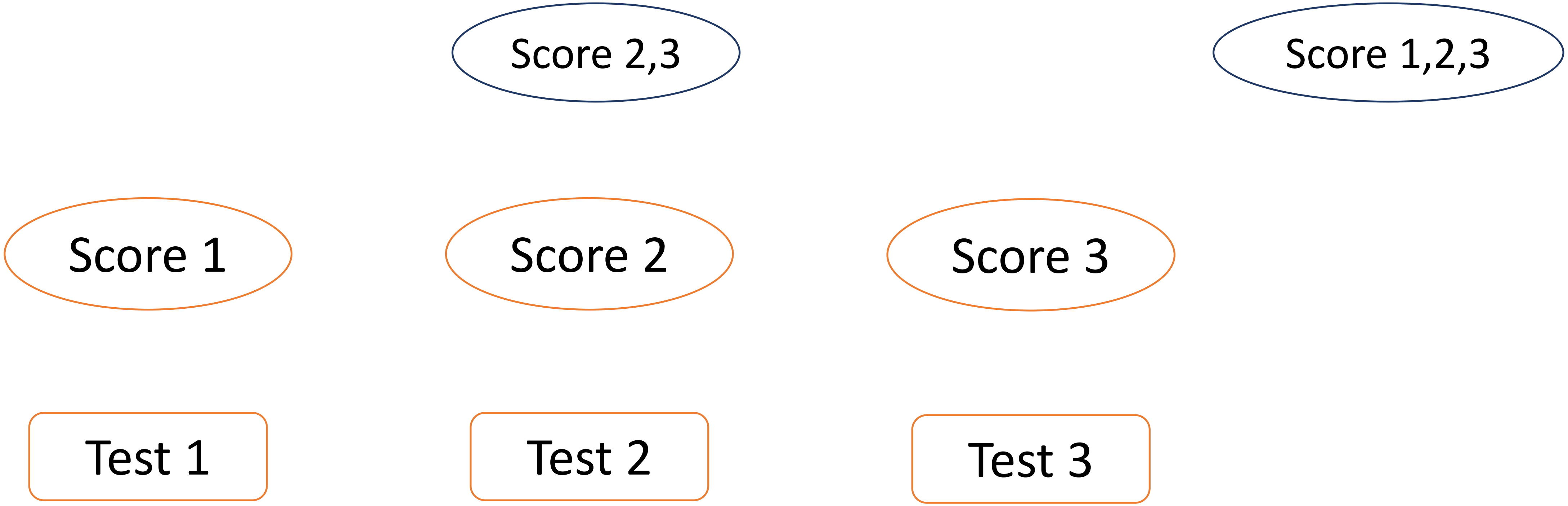
Score 2

Score 3

Test 1

Test 2

Test 3



Score 2,3

Score 1,2,3

Score 1

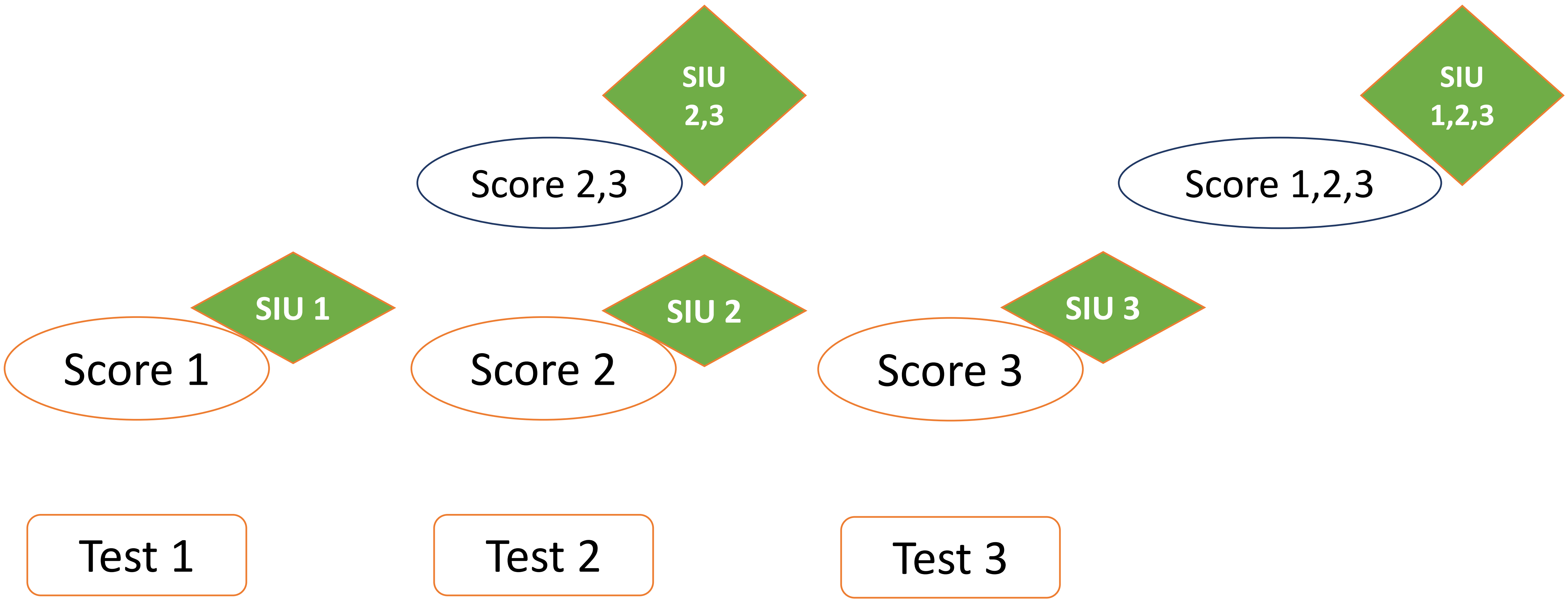
Score 2

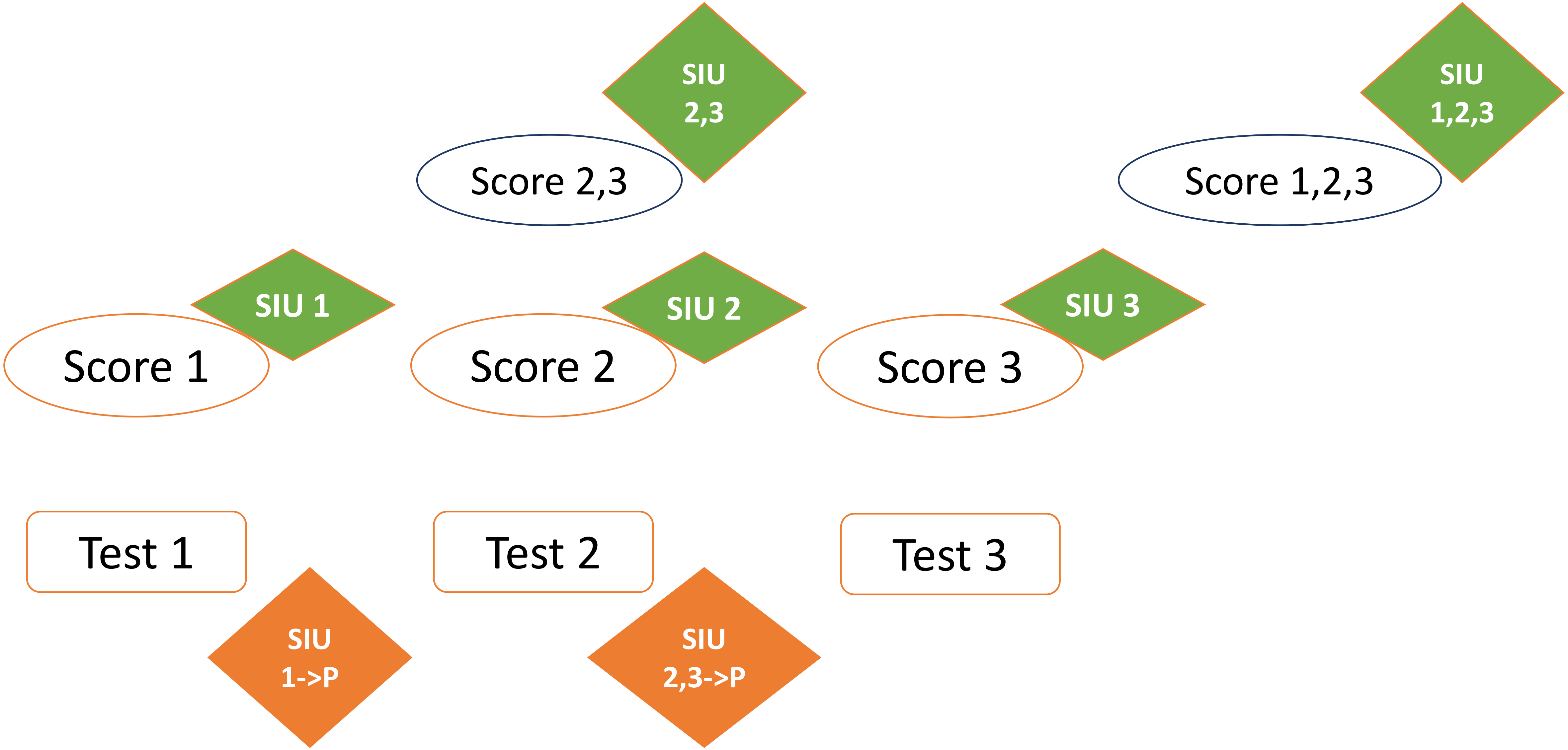
Score 3

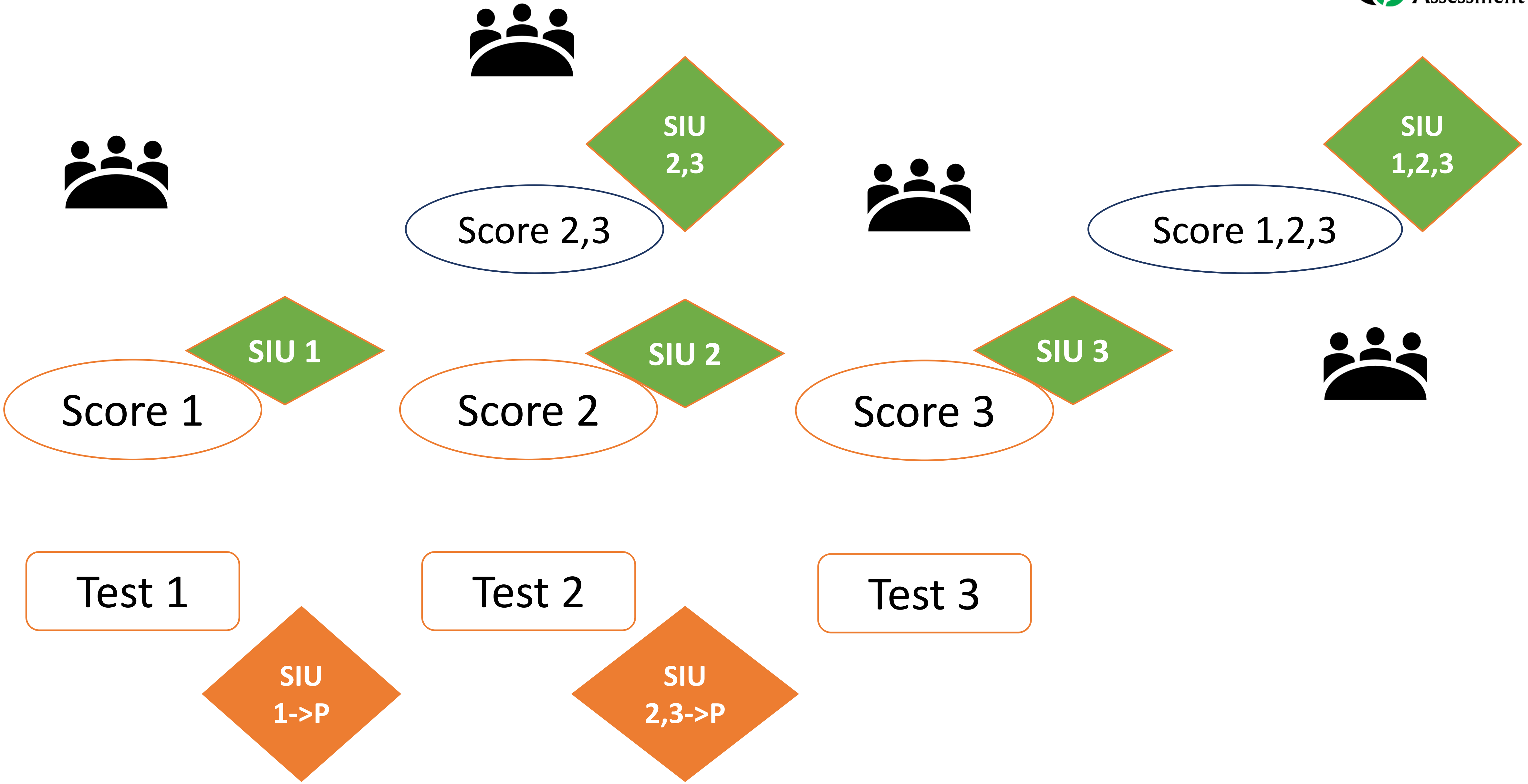
Test 1

Test 2

Test 3


SIU 2,3
Score 2,3
SIU 1,2,3
Score 1,2,3
SIU 1
Score 1
SIU 2
Score 2
SIU 3
Score 3
Test 1
Test 2
Test 3


SIU 2,3
Score 2,3
SIU 1,2,3
Score 1,2,3
SIU 1
Score 1
SIU 2
Score 2
SIU 3
Score 3
Test 1
Test 2
Test 3
SIU 1->P
SIU 2,3->P



SIU 2,3

Score 2,3

SIU 1,2,3

Score 1,2,3

SIU 1

Score 1

SIU 2

Score 2

SIU 3

Score 3

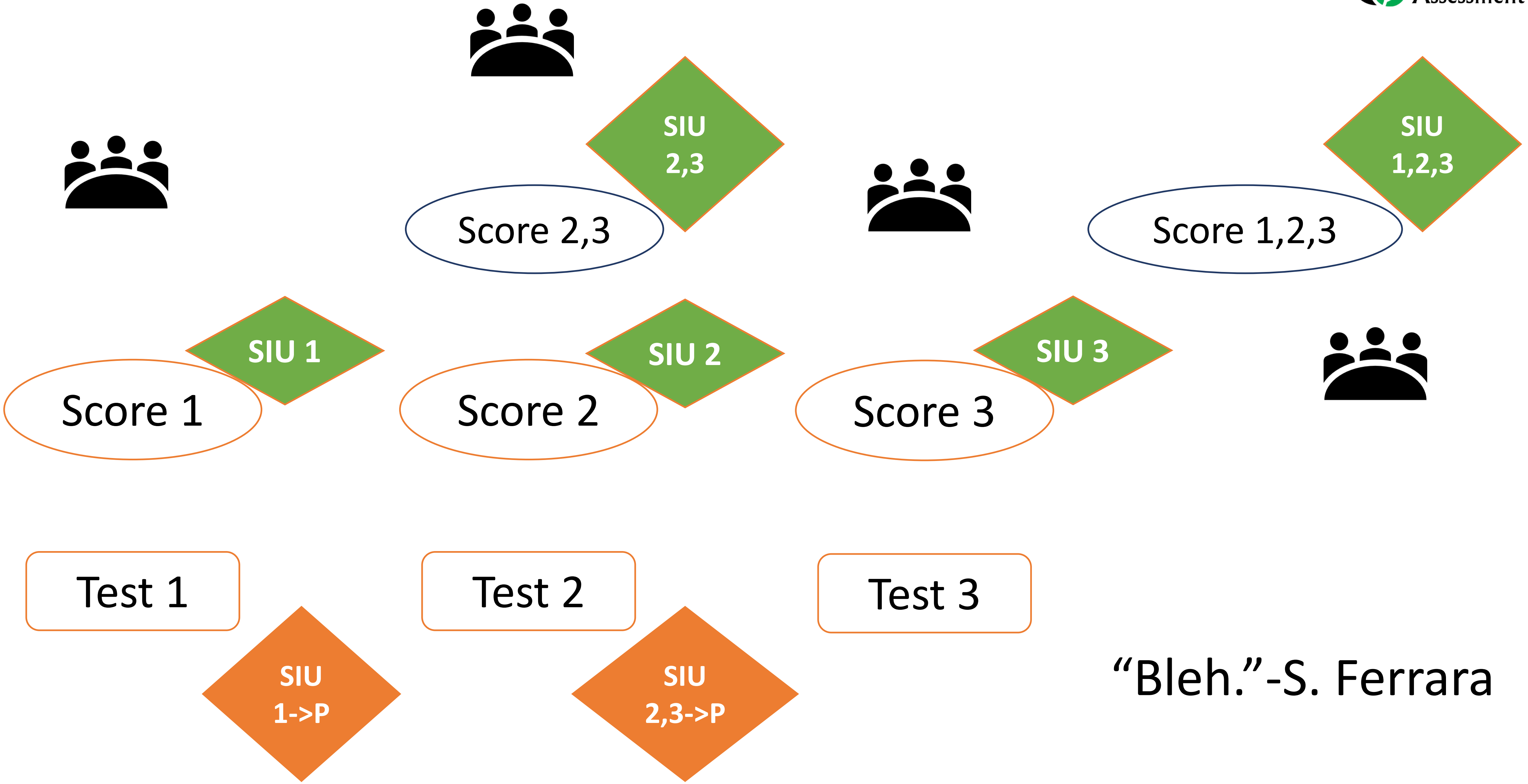
Test 1

Test 2

Test 3

SIU 1->P

SIU 2,3->P



SIU 2,3

Score 2,3

SIU 1,2,3

Score 1,2,3

SIU 1

Score 1

SIU 2

Score 2

SIU 3

Score 3

Test 1

Test 2

Test 3

SIU 1->P

SIU 2,3->P

“Bleh.”-S. Ferrara

Standard Setting and Through-Year: Open Questions

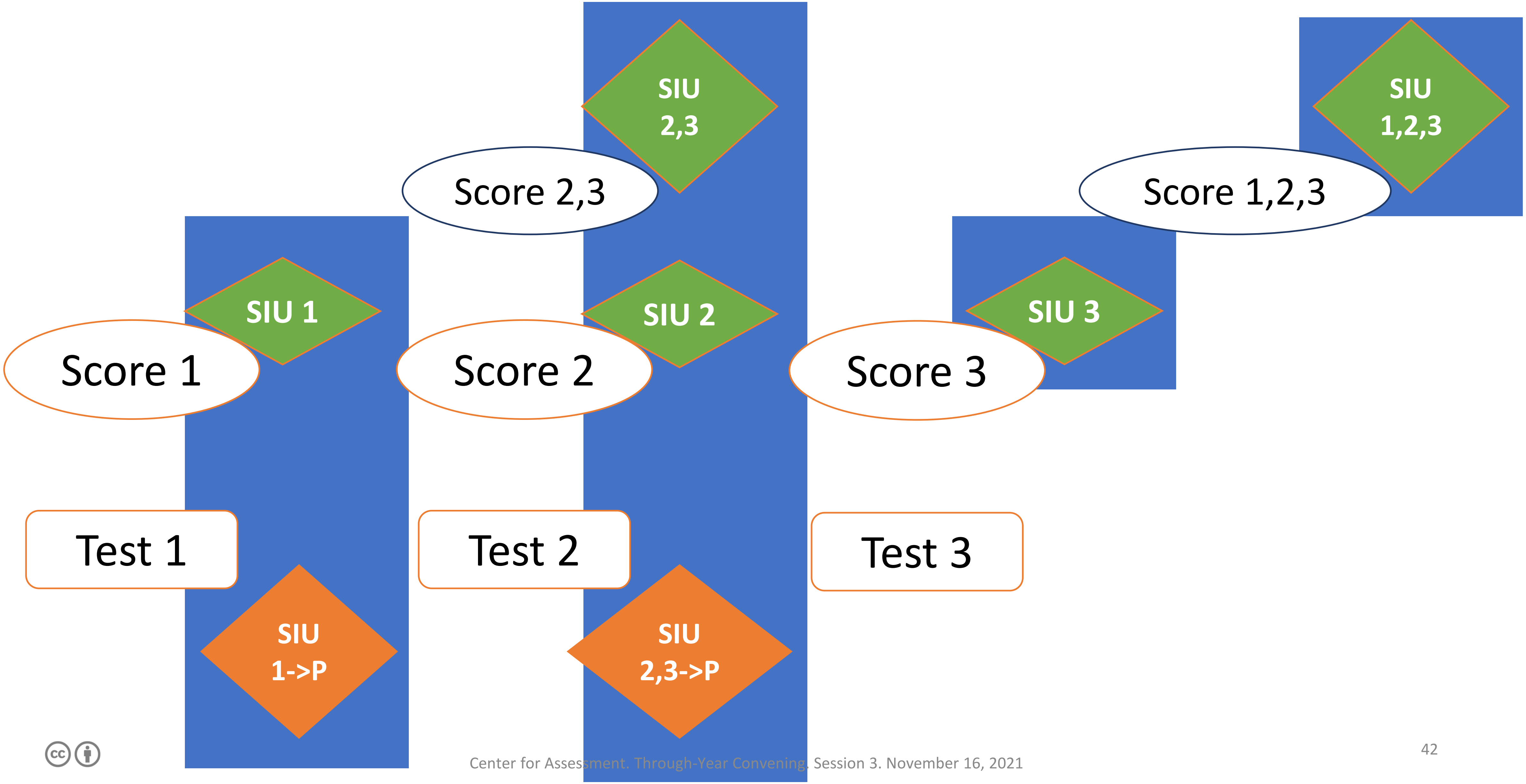
- **Expectations, Part 1.** How many sets of expectations will the through-year program need? If multiple, how will they relate to each other?
- **Expectations, Part 2.** “With summative assessment we are often looking for ‘mastery’ or ‘achievement’ of the standards. What are we looking for at each assessment point?” – S. Davis-Becker
- **Relation to Aggregation.** How does standard setting approach depend on aggregation method?

5. Reporting

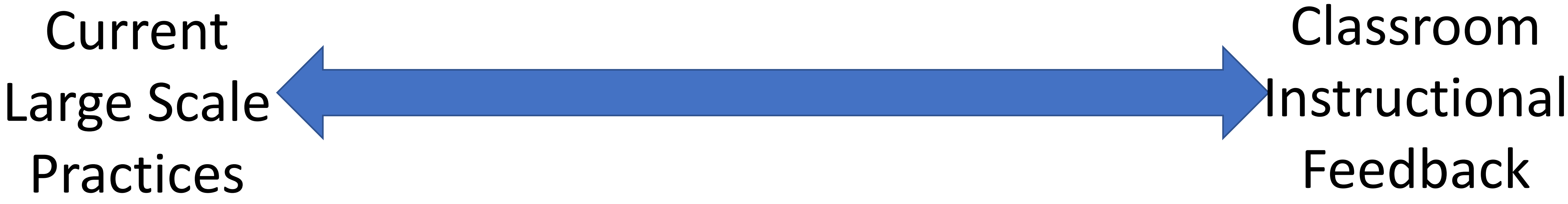
Common Issues In Reporting

- Defining:
  - Who the report is about
  - Who the report is for
  - What they are meant to do with the report
- Reporting Metrics
  - Status (typically in terms of scale scores and classifications)
  - Change
- Comparisons
- Timeliness
- Infrastructure

E.g., Zenisky (2019)


SIU
2,3
Score 2,3
SIU
1,2,3
Score 1,2,3
SIU 1
Score 1
SIU 2
Score 2
SIU 3
Score 3
Test 1
Test 2
Test 3
SIU
1->P
SIU
2,3->P

Considering Reporting & Scoring



Current Large Scale Practices

Classroom Instructional Feedback

- Multi-Month Turn Around (strong QC)
- Secured
- Highly Reliable
- Clearly delimited interpretation

- Rapid Turn Around
- Open
- Fine Grained Information
- Tailored to meet instruction

Open Questions: Reporting

- What information do we need to report, and when?
  - Will that information be useful (i.e., support the theory of action)
- How do we handle the **complexity** that some additional uses may require (e.g., a set of highly contextualized, interconnected reports)?
- For instructional uses, what kinds of **additional supports** (including recommended actions or supplemental connections to curriculum) are be needed?

2. Invited Presentations

Welcome to Our Panelists

Discussion

- We asked each of our panelists to discuss a limited number of technical and/or logistical issues that keep them up at night.
- What’s the issue(s) and how are you approaching it?

Through Year Assessment:

Promises and Challenges

Ye Tong, Ph.D.

Pearson

November 16, 2021

Pearson

General Solution

- 2 + 1 Through Year Strategy
- The 2 are interim exams and can feed into actional information for instruction
- The 1 is mini summative and can feed into accountability by itself
- Performance on the 2 interims can help start the mini summative
- Monitor within-grade and cross-grade growth

General Solution

- Do not combine scores across interims and summative
- Careful with within-grade growth usage and its implications
- Coherent and balanced system (formative)

Challenges

- Alignment, scope and sequence
- Scores and interpretations
- Accommodations and accessibility

Alignment, Scope and Sequence

- Alignment challenges with typical product
  - Customization?
- Scope and sequence at local level
  - Opportunity to learn
  - Implication on within-grade growth
- Cumulative intelligent blueprint, Transcend
  - State can determine overall blueprint
  - Allow local selection of standards for interim
  - Cumulative in nature

Scores and Interpretations

- Do not recommend combination of scores across segments
  - Exception example: social studies?
- Need to have all scores on the same scale
  - IRT models
  - How to establish link
  - When to field test
- Inferences on standards not assessed
- Vertical scale and within grade growth
- Off grade content/floor and ceiling effect

Accessibility and Accommodations

- Interim product versus state summative
  - Translation and transadaptation
  - Creation of embedded American Sign Language Video content
  - Support of screen readers and refreshable braille displays
  - Paper base delivery for students requiring a paper accommodation
  - Printed braille
- Student experience and equity consideration

nwea State Solutions

Threading the Needle with the Integrated Through-Year System

One coherent system to drive action

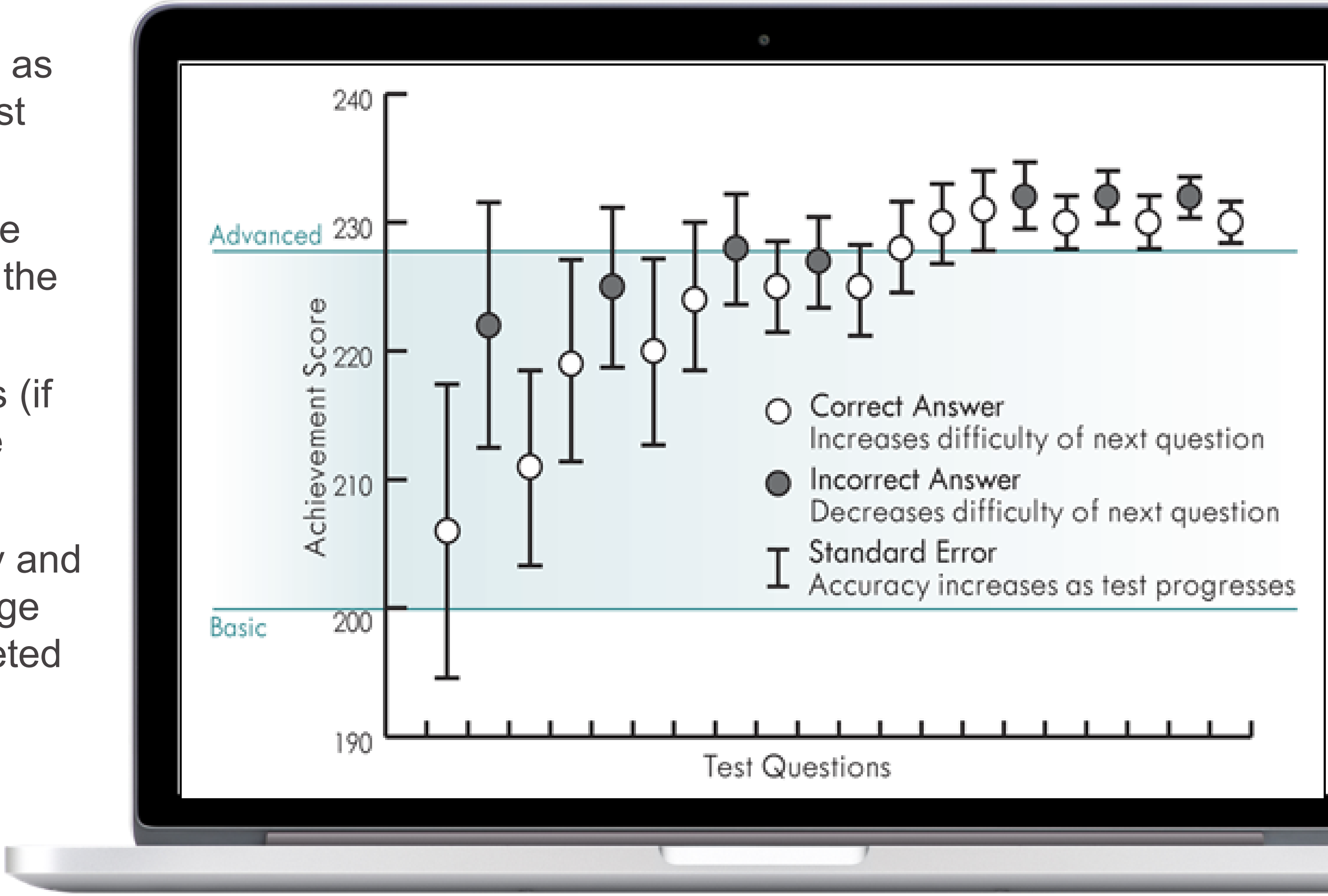
Garron Gianopulos, Learning and Assessment Engineer

Technical Questions

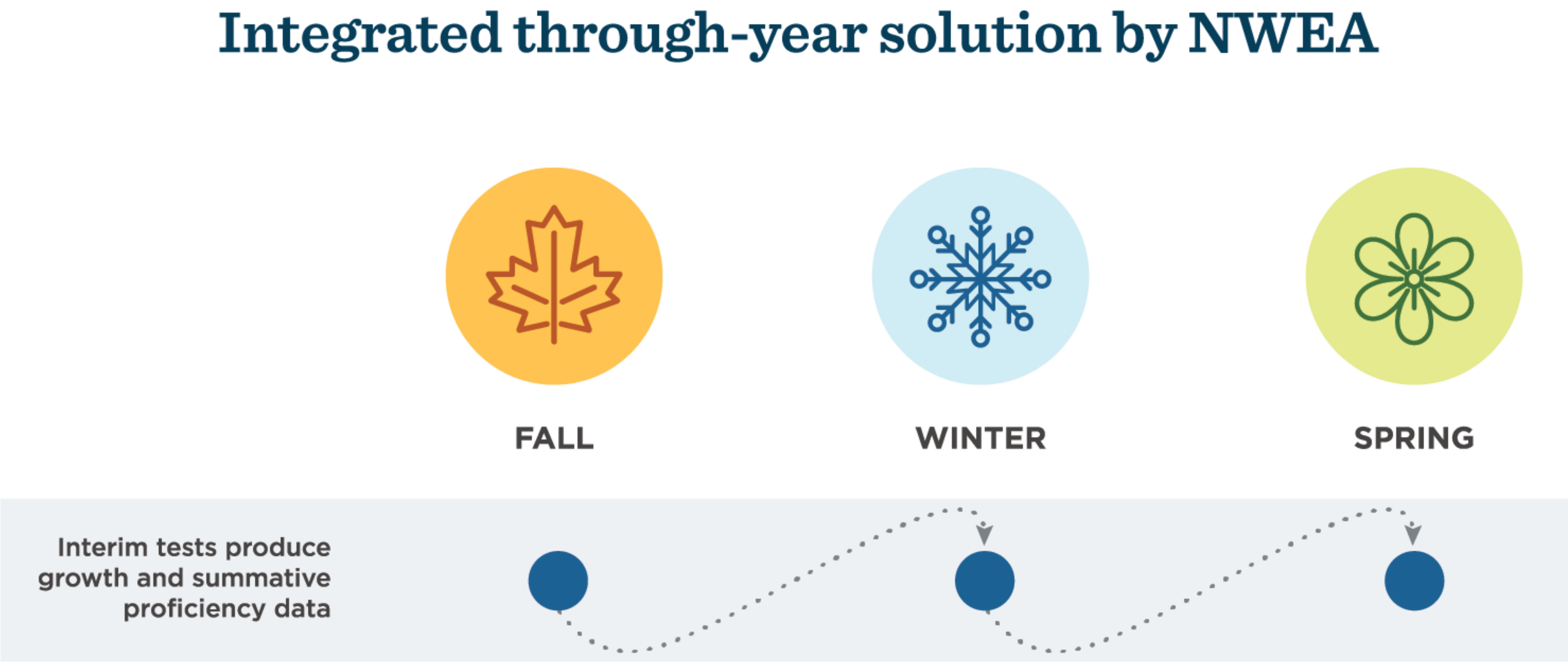
- What are the implications of your program's score interpretive claim for the timing of field testing?
- As a practical matter, how will your program achieve the volume of field testing necessary to sustain it?

Why item-level computer adaptive tests (CATs)?

- CATs can reduce test length by as much as 50% when compared to non-adaptive test forms (Weiss, 1982).
- CATs can improve efficiency (produce the most information about student ability in the fewest possible items).
- CATs can reduce ceiling and floor effects (if there are enough items at the tails of the ability distribution).
- While CATs promise increased efficiency and improved measurement, they require large calibrated item pools aligned to the targeted population distribution to realize these benefits.



High-level Overview of Integrated Through-Year System



Integrated through-year solution by NWEA

Interim tests produce growth and summative proficiency data

Instructional feedback -- Within-year Growth -- Summative Determinations

Timing

- What are the implications of your program's score interpretive claim for the timing of field testing?
  - Within-year growth interpretations require item parameter invariance across seasons; therefore, items should be calibrated in multiple time points to check the assumption of item parameter invariance.
  - Spring summative determinations require representative samples of students from the targeted population at the end of the year after opportunity to learn is complete; therefore, the summative scale should reference the spring test event.
  - All other field tested items should be placed onto the same summative scale.

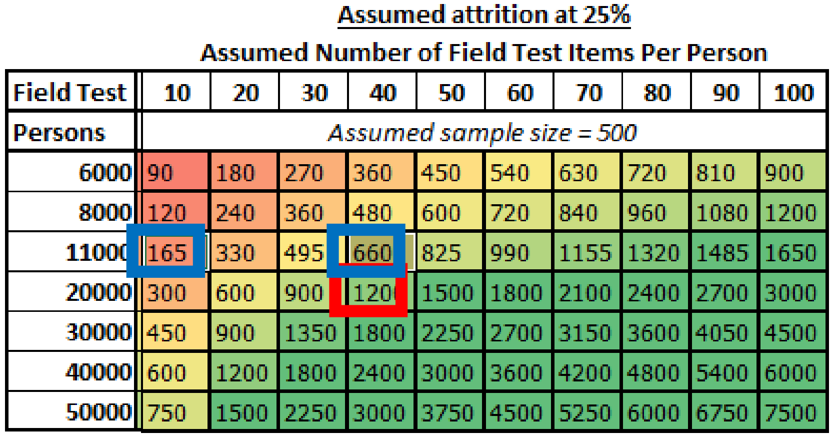
How large an item pool?

- A conventional rule of thumb is that an item pool should have enough items to construct 5–10 test forms (Parshall, Spray, Kalohn, & Davey, 2002; Stocking, 1994).
- Needed items = test length (40) × number of admins. (4) × 10 forms[1]
- 800 to 1,600 items are needed
- CAT simulations using mixed integer linear programming thus far suggest minimally 800 items are needed.[2]
- Need to reserve most informative items for proficiency classifications in spring.

1. 1 for each season (3) plus 1 breach. 2. Minimum needed may be less that 800 using our constraint engine which uses quadrature programming and longitudinal item exposure procedures, rather than MILP.

As a practical matter, how will your program achieve the volume of field testing necessary to sustain it?

- A single stand-alone field test (SAFT) of 40 items with 20,000 students should yield >1200 calibrated items (Rasch model), OR
- An initial SAFT of 40 items with 11,000 students plus a 2nd test with 10 embedded field test items should yield > 800 calibrated items.
- Ongoing embedded field testing across each season to continually replenish the item pool.
- Under R&D: Optimal design of experiments (Lu, 2014), automated item generation, and priors based on item difficulty models (Gianopulos & Kim, 2019).



Assumed attrition at 25%

Assumed Number of Field Test Items Per Person

Field Test	10	20	30	40	50	60	70	80	90	100
Persons	*Assumed sample size = 500*									
6000	90	180	270	360	450	540	630	720	810	900
8000	120	240	360	480	600	720	840	960	1080	1200
11000	165	330	495	660	825	990	1155	1320	1485	1650
20000	300	600	900	1200	1500	1800	2100	2400	2700	3000
30000	450	900	1350	1800	2250	2700	3150	3600	4050	4500
40000	600	1200	1800	2400	3000	3600	4200	4800	5400	6000
50000	750	1500	2250	3000	3750	4500	5250	6000	6750	7500

Conclusion

- Large calibrated item pools aligned to the targeted population distribution are necessary to realize the expected benefits of CAT, including reduced test length, improve measurement efficiency, and reduced floor/ceiling effects (Weiss, 1982).
- Field testing for three CATs per year requires a larger initial effort and ongoing effort to maintain the item pool.
- Items most useful to proficiency classification can be reserved for the end-of-year CAT to maximize classification accuracy of summative determinations.
- Item difficulty modeling can also be used to generate item variants targeted at achievement level cut scores. In conjunction with optimal design of experiments, item pools can be replenished more efficiently.

References

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Davidson, A. H. (2020, February 28). *Alignment of MAP Growth items and the Georgia Standards of Excellence*. EdMetric report provided to NWEA.

Davey, T., Pitoniak, M. J., & Slater, S. C. (2015). Designing computerized adaptive tests. In *Handbook of test development* (pp. 483–500). Routledge.

Ferrara, S., & Lewis, D. (2012). The item-descriptor (ID) matching method. In G. Cizek (Ed.), *Setting Performance Standards* (2nd ed.) (pp. 255–282). New York: Routledge.

Gianopulos, G. & Kim, J. (2021, June). *Integrating item difficulty modeling into test design for continual improvement*. In G. Gianopulos (Chair), *The past, present, and future of item difficulty modeling*. Symposium conducted at the annual Meeting of the National Council on Measurement in Education, Baltimore, MD.

Lu H-Y (2014) Application of Optimal Designs to Item Calibration. PLoS ONE 9(9): e106747. https://doi.org/10.1371/journal.pone.0106747

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer Science & Business Media.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report 94-5). Princeton, NJ: Educational Testing Service.

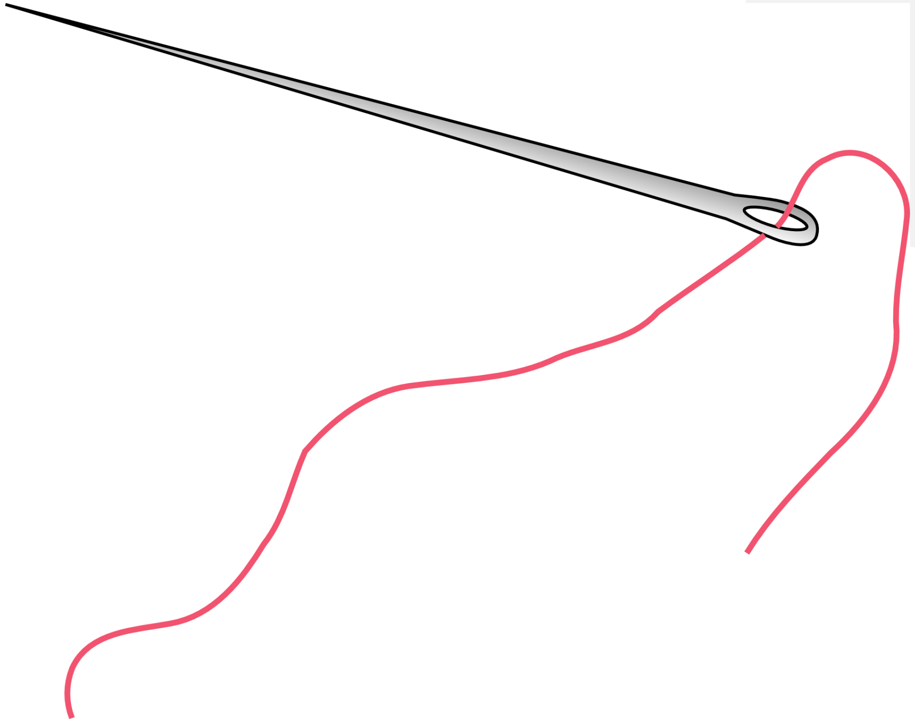
Van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics, 31*(1), 81–99.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, *6*(4), 473-492.

2. Invited Presentations

Center for Assessment

Closing



- Now that we’ve increased your anxiety—sorry—our next session in **30 minutes** will focus on threading the needle to see if we can find ways to make through-year assessment systems work to improve learning opportunities for students.