OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

SCHOOL SUPPORT AND ACCOUNTABILITY

# 2023 State Assessment Conference
# 1A: Test Design and Development

September 26, 2023

# FOCUS AREA: ASSESSMENT PEER REVIEW

## 1A – TEST DESIGN AND DEVELOPMENT

Meagan Karvonen, ATLAS, karvonen@ku.edu
Nathan Dadey, Center for Assessment, ndadey@nciea.org
Brooke Nash, ATLAS, bnash@ku.edu
Chris Rozunick, Texas Education Agency,
Christine.Rozunick@tea.texas.gov

# A note About this Conference/Session

- The purpose of this conference/session is to provide an opportunity for State education agency (SEA) staff to interact and engage with relevant experts and other SEA staff about the Department's assessment peer review.
- The observations and opinions of the session presenters are their own.

# Session Overview

- Framing and overview
- Issues and options
- State examples
- Responding to peer review requirements
- Q&A

# Framing & Overview


5

# Which Critical Elements?

Primarily addressed in:

2.1: Test design and development
2.2: Item development
3.1: Overall validity, validity based on content

# 2.1: Test Design and Development

The State's test design and test development process is well-suited for the content, is technically sound, **aligns the assessments to (1) the depth and breadth of the State's academic content standards** for the grade that is being assessed;

- the **depth and breadth** of the State's grade-level academic content standards in terms of **balance of content** (i.e., knowledge, cognitive process, cognitive complexity).

# More on Critical Element 2.1

- Statement of purposes, intended uses
- Blueprints support test development – depth and breadth
- Assessment tailored to knowledge and skills in the standards, include appropriately complex applications
- Computer Adaptive Testing (CAT) item pool, selection procedures support test design

# 2.2: Item Development

The State uses reasonable and technically sound procedures to develop and select items to:

- Assess student achievement based on the State's academic content standards in terms of content and cognitive process, including higher-order thinking skills.

# 3.1: Overall Validity + Content (1)

The State has documented adequate overall validity evidence for its assessments consistent with nationally recognized professional and technical testing standards.

# 3.1: Overall Validity + Content (2)

Assessments measure the knowledge and skills specified in the State's academic content standards:

- adequate alignment between assessments and the academic content standards the assessments are designed to measure in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity;
- assessments address the depth and breadth of the content standards;

(Not in this session: For AA-AAAS, alignment of assessments to content standards)

# Other Critical Elements

May also impact:

4.2: Fairness + Accessibility

4.6: Multiple Versions

4.7: Technical Analysis, Ongoing Maintenance

5.3: Accommodations

Indirect impacts: 4.1 (Reliability), 4.4 (Scoring), 6.4 (Reporting)

# Caveats

- This session's examples will be based on variants of through-year designs
  - Can discuss other designs during Q&A

- Alignment is covered in session 1B

# Through-Year (TY) Assessment

A through-year assessment program is one that is

- Administered in multiple distinct sessions during a school year, and
- Intended to support (a) the production and use of a summative determination, and (b) one or more additional aim(s).

**“Full Domain” Designs: Each Assessment Covers the Full* Standards**

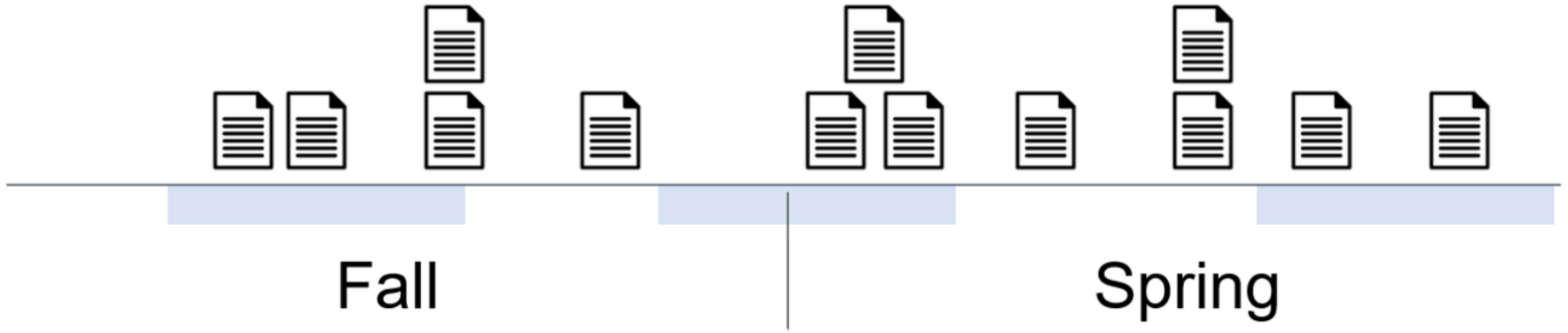

**“Modular” Design: Each Assessment Covers A Small Group of Standards**


**“Modular” Design: Each Assessment Covers A Single Standard**

**Texas Through-Year Assessment Pilot**


**"Modular" Design: Each Assessment Covers A Small Group of Standards**


**Dynamic Learning Maps Instructionally Embedded Assessments**



# Multiple Approaches Handout

Download for reference:

- Summary of assessment programs featured in Focus 1 sessions
- Focus 1 slides (excerpt from plenary session)

# Issues and Options



18

# Topics

- Test Design
  - Purposes and intended interpretations and uses
  - Test blueprint
  - Adaptive administration procedures
- Test Development
  - Item development procedures
  - Item selection procedures

# Test Design - Purpose & Use (1)

Some common issues:

- Too many and/or unclear purposes and uses as they relate to the academic content standards.
  - Including intended interpretation of performance expectations.
- Purposes and uses that don't align well with the goals of the new/different assessment approach.
- Test design choices are made that don't align well with purposes and uses.

# Test Design - Purpose & Use (2)

Suggestions:

- Use a theory of action (or other type of logic model) to identify the goals and long-term outcomes of the new/different assessment approach.
- Articulate clear statements of intended interpretations and uses of assessment results. **Get agreement first and then design!**

# Test Design - Purpose & Use (3)

**Suggestions:**

- Focus on connections between intended uses and each component of test design, including:
  - purposes outside the purview of peer review but still impact design.
  - working iteratively from results reporting designs (e.g., score reports, reporting dashboards).
- If possible, conduct small-scale pilot study to evaluate prior to full scale development (does the design meet intended uses).

# Test Design - Blueprint (1)

Some common issues:

- Lack of understanding of the domain or content structure can lead to unnecessary blueprint requirements and misunderstanding of how best to achieve depth and breadth of content standards.
- Depth and breadth of content standards, at each level of intended inference:
  - within individual forms but across schools, districts and/or years
  - within and across multiple administrations within the same year
    - Local pacing and Opportunity to Learn (OTL) considerations.

# Test Design - Blueprint (2)

Some common issues:

- Release/reuse requirements that will impact blueprint
  - Can items/tests be reused?
  - Is there any requirements around release?
- Number of administrations and test length
  - How short is too short?
  - What are the trade-offs?
    - Reduced or no subscore reporting for within-year assessments.

# Test Design - Blueprint (3)

Suggestions:

- Clear description of content structure and how it relates to the blueprint specifications.
- Determine (early on) which parts of which assessments will be used in summative calculations.
  - Blueprints can specify depth and breadth of content standards for the “total assessment” (i.e., for summative reporting), while still supporting other intended uses (e.g., reporting student skill mastery throughout the year).

# Test Design - Administration Considerations (1)

Some common issues:

- If using computer adaptive administration, adaptive algorithms and procedures used for typical item-level or even multi-stage CATs may need to be adjusted for:
  - assessments delivered several times a year or “year-round” (e.g., imposing limits so that items that appear in prior admins do not appear in later admins).
  - assessments scored using psychometric models other than IRT.
  - assessments that might include additional assignment rules or methods (e.g., teacher selected content).

# Test Design - Administration Considerations (2)

Suggestions:

- Administration procedures and algorithms should be clearly defined, tested, and continuously monitored to ensure blueprint coverage is met at each level of inference intended by the design.
- Work closely and early on with technology teams on requirements needed to support adaptive administration of multiple approaches to assessment.

# Test Development - Item Development Procedures (1)

Some common issues:

- Size and coverage of item pools needed to support multiple approaches to assessment.
  - Existing banks may not align well to the state content standards or other state specific style guidelines, etc.
  - Existing banks may not meet the intended specifications of design (e.g., an item written for a test scored using IRT may not work well for a test scored using diagnostic classification modeling (DCM)).
  - Item writer training and procedures may need to be adjusted to ensure items are developed that meet intended uses

# Test Development - Item Development Procedures (2)

Suggestions:

- If using an existing item bank, “pressure test” the content against design criteria. Do this early on to ensure it will meet the needs!
- Consider using evidence-centered design task models to support item writing.

# Test Development - Item Selection Procedures (1)

Some common issues:

- Mismatch between item selection procedures intended to meet test form requirements and the items available in the bank.

# Test Development - Item Selection Procedures (2)

Suggestions:

- Define item selection requirements first, then evaluate item bank and item development needs based on requirements and understanding of item selection procedures that will be implemented.

# State Examples



32

# Texas Through-Year Assessment Pilot: Purpose and Use

- State content aligned, valid, and reliable assessment system that could replace existing assessments.
- Progress monitoring system that provides timely data and information to support instruction.
- Assessments that are minimally disruptive to instructional time.

# Approach to Test Design (1)

- Since our number one goal is replace STAAR, we began each design consideration with reviewing current practice along with all possible options.
- Prioritized timing as well in design considerations.
- First two of three opportunities each year need to be shorter than the third.

# Approach to Test Design (2)

- Working in targeted grades across all levels and subjects across Math, Science, and Social Studies - we are judging feasibility not moving to operational implementation.
- Planning to begin RLA pilot of a couple grades next year.

# Intended Score Report Information

- Score Reporting is still a work in progress.
- We have a small set of reports for the current pilot but we are still taking pilot participant feedback into account before finalizing reports.
- We currently have mainly static reports but they are available online. There is also dynamic data available to educators, parents, and the public.

# Trade Offs

- We are exploring cumulative scoring that takes into account student proficiency demonstrated throughout the year - this requires research and policy considerations.
- Data literacy of educators interacting with the pilot has been a huge need.
- This is a pilot and everyone likes it. But it doesn’t count yet.

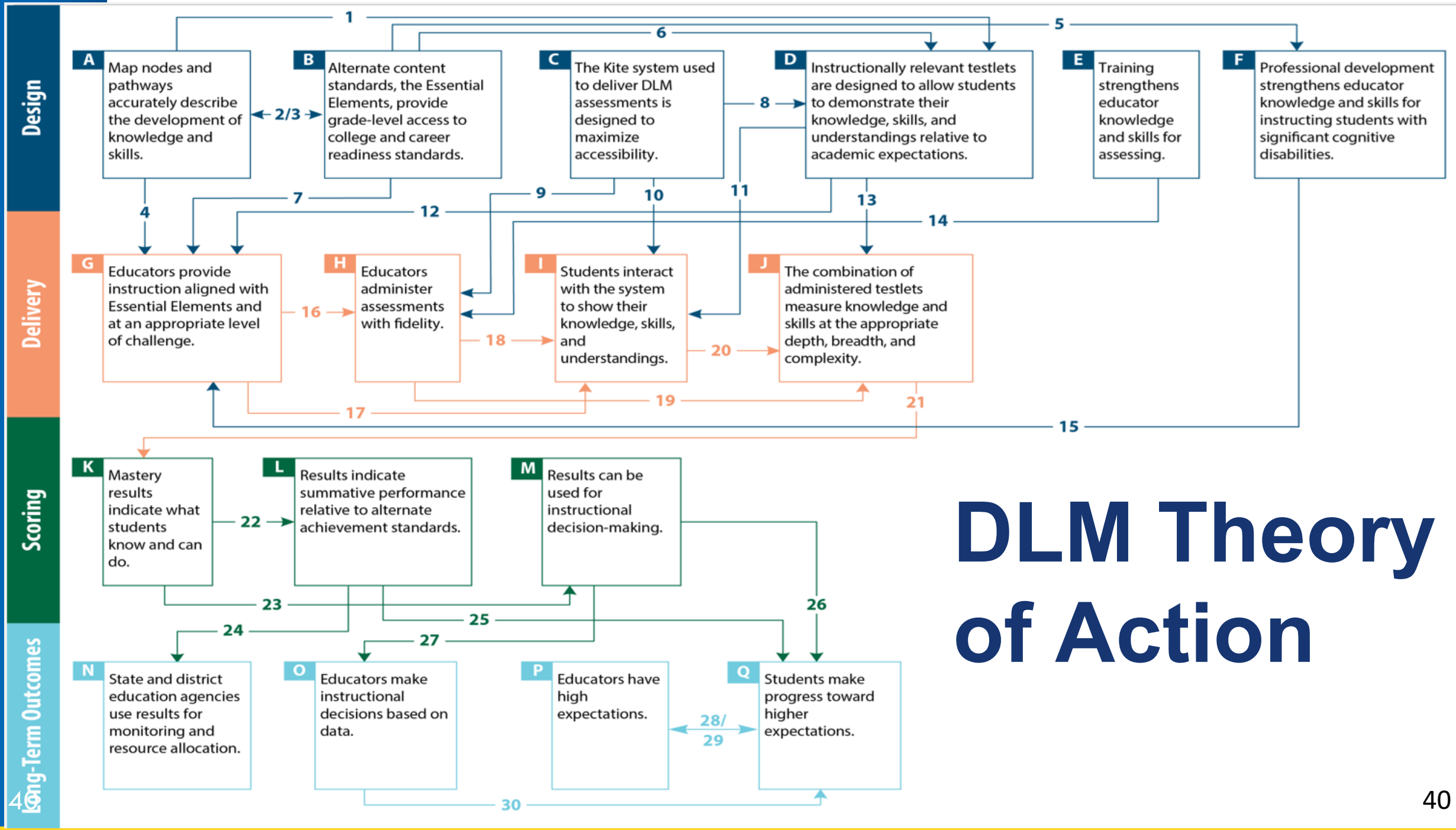
# Dynamic Learning Maps (DLM) - Instructionally Embedded Assessment Model

- Assessments in English language arts and mathematics for have been operationally administered in several states since 2014-2015.
- The DLM alternate assessment system serves students with significant cognitive disabilities in grades 3-8 and high school.
- Results are intended to support interpretations about what students know and are able to do in each assessed content area.
- Results provide information that can be used to guide instructional decisions as well as information appropriate for use with state accountability programs.

# Some Features of the DLM Instructionally Embedded Assessment Model

- Based on learning maps that describe how students acquire knowledge and skills.
  - The learning maps provide a framework that supports inferences about student learning needs.
- A set of learning targets for instruction and assessment aligned to grade-level academic content standards.
- Instructionally relevant assessments
- Accessibility by design
- Assessment results that are readily actionable

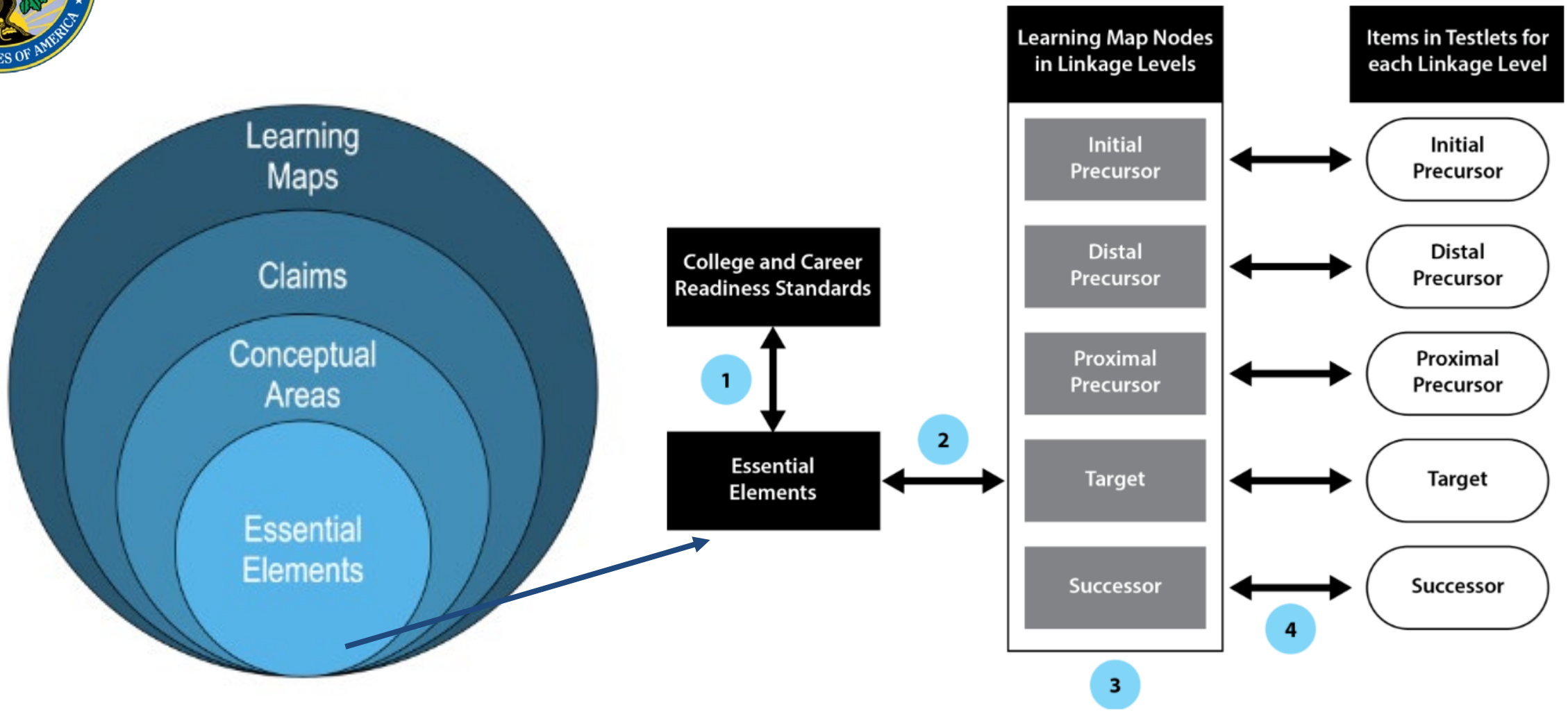
Met all peer review requirements for use as an accountability assessment!




# DLM Theory of Action


40

# DLM Content Structures



# Example Blueprint

| Claim | Conceptual Area | Essential Element | Description |
|---|---|---|---|
| 1 | **Students demonstrate increasingly complex understanding of number sense. Requirement: Choose three Essential Elements from Claim 1 in at least two different conceptual areas.** | | |
| | **M.C1.1** | M.EE.5.NF.1 | Identify models of halves (1/2, 2/2) and fourths (1/4, 2/4, 3/4, 4/4). |
| | | M.EE.5.NF.2 | Identify models of thirds (1/3, 2/3, 3/3) and tenths (1/10, 2/10, 3/10, 4/10, 5/10, 6/10, 7/10, 8/10, 9/10, 10/10). |
| | **M.C1.2** | M.EE.5.NBT.1 | Compare numbers up to 99 using base ten models. |
| | | M.EE.5.NBT.3 | Compare whole numbers up to 100 using symbols (<, >, =). |
| | | M.EE.5.NBT.4 | Round two-digit whole numbers to the nearest 10 from 0—90. |
| | **M.C1.3** | M.EE.5.NBT.5 | Multiply whole numbers up to 5 x 5. |
| | | M.EE.5.NBT.6-7 | Illustrate the concept of division using fair and equal shares. |
| 2 | **Students demonstrate increasingly complex spatial reasoning and understanding of geometric principles. Requirement: Choose one Essential Element from Claim 2.** | | |
| | **M.C2.1** | M.EE.5.G.1-4 | Sort two-dimensional figures and identify the attributes (angles, number of sides, corners, color) they have in common. |
| | | M.EE.5.MD.3 | Identify common three-dimensional shapes. |
| | **M.C2.2** | M.EE.5.MD.4-5 | Determine the volume of a rectangular prism by counting units of measure (unit cubes). |
| 3 | **Students demonstrate increasingly complex understanding of measurement, data, and analytic procedures. Requirement: Choose two Essential Elements from Claim 3 in different conceptual areas.** | | |
| | **M.C3.1** | M.EE.5.MD.1.a | Tell time using an analog or digital clock to the half or quarter hour. |
| | | M.EE.5.MD.1.b | Use standard units to measure weight and length of objects. |
| | | M.EE.5.MD.1.c | Indicate relative value of collections of coins. |
| | **M.C3.2** | M.EE.5.MD.2 | Represent and interpret data on a picture graph, line plot, or bar graph. |
| 4 | **Students solve increasingly complex mathematical problems, making productive use of algebra and functions. Requirement: All students are assessed on the Essential Element in Claim 4.** | | |
| | **M.C4.2** | M.EE.5.OA.3 | Identify and extend numerical patterns. |

# Test Development

- Testlets are based on nodes for one linkage level of one EE.
- Each testlet contains three to nine items.
- All testlets begin with a nonscored engagement activity.
- Within testlets, several item types are used in DLM testlets:
  - Multiple-choice single-select.
  - Multiple-choice multiple-select.
  - Select text (ELA only).
  - Matching lines (mathematics only).

# Test Development Principles

- The DLM System uses evidence centered design (ECD) procedures to develop test specifications and task templates for item creation that also incorporate UDL principles (Bechard et al., 2019).
  - The ECD approach is structured as a sequence of test development layers that include (a) domain analysis, (b) domain modeling, (c) conceptual assessment framework development, (d) assessment implementation, and (e) assessment delivery (Mislevy & Riconscente, 2005).
  - Incorporating principles of UDL allows students to respond to items free of barriers while emphasizing accessibility and offering multiple ways to demonstrate skills.

**Claim:** ELA.C1 Students can comprehend text in increasingly complex ways.
**Conceptual Area:** ELA.C1.2 Construct Understandings of Text
**General Education Content Standard:** ELA.RI.6.2 Determine a central idea of a text and how it is conveyed through particular details; provide a summary of the text distinct from personal opinions or judgments.
**Essential Element:** ELA.EE.RI.6.2 Determine the main idea of a passage and details or facts related to it.

DYNAMIC® LEARNING MAPS

**Essential Questions**
- Can the student identify the main idea of a passage?
- Does the student recognize that details and facts can relate to the main idea of a passage?

| Vocabulary | (a) Initial Precursor | (b) Distal Precursor | (c) Proximal Precursor | (d) Target | (e) Successor |
|---|---|---|---|---|---|
| *Concepts* | environment, object and person identification, object/picture association | concrete detail identification | detail identification, key details | main idea, detail identification, main idea/detail association | central idea, key details, key detail/central idea association |
| *Words* | naming words (dog, ball, girl, etc.), wh words (who, what, which, where) | find, wh words (who, which, what, where, when) | find, wh words (who, which, what, where, when), detail | main idea, details, wh words (who, which, what, where, when) | important, detail, support, wh words (who, which, what, where, when), how, main idea |

| (a) Initial Precursor Nodes | Node Descriptions | Node Observations | # Items |
|---|---|---|---|
| F-154-Can demonstrate understanding of property words corresponding to the objects used during familiar routines | Can demonstrate a receptive understanding of the property words that describe the objects that accompany familiar games or routines. | During a shared reading activity with the student, the student is able to identify items based on their property descriptions. | 3-5<br>☒ TA<br>☐ Blind/VI (B) |

**(a) Questions to Ask**
- Does the student recognize property words?
- Show me the (property word) one.

**(a) Misconceptions**
- The student indicates a different object.
- The student indicates multiple objects.
- The student attends to other stimuli.
- The student does not respond.

| (b) Distal Precursor Node | Node Description | Node Observation | # Items |
|---|---|---|---|
| ELA-1141 Can identify concrete details in familiar informational texts | Can identify the concrete details, such as individuals, events, or ideas in familiar informational texts. | When asked to recall a concrete detail from a familiar informational text, the student is able to identify the correct detail from the text. | 5<br>☐ TA<br>☐ Blind/VI (B) |

**(b) Questions to Ask**
- Does the student recognize that informational texts contain concrete details?
- Can the student identify the correct detail to answer a question?
- Who is John?

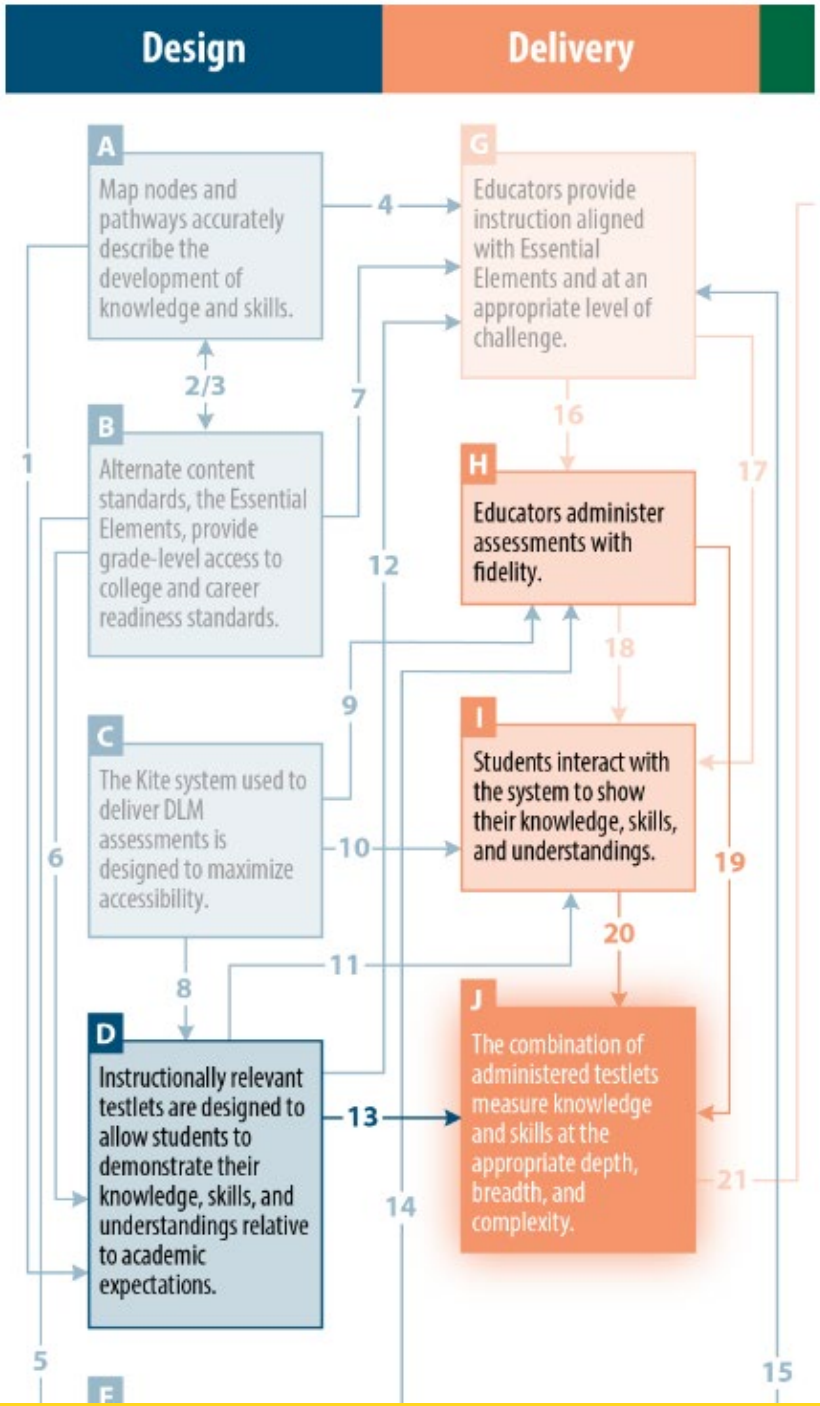
**(b) Misconceptions**
- The student chooses a detail unrelated to the particular question.
- The student attempts to use the illustration to answer a question about a concrete detail rather than the text

| (c) Proximal Precursor Node | Node Description | Node Observation | # Items |
|---|---|---|---|
| ELA-1462 Can identify the key details in a paragraph of an informational text | Can determine which details in a paragraph of an informational text are important. | After reading an informational text, the student can identify each of the key details in the paragraph. | 5 |

Excerpt of Task Model Template

# Testlet Assignment Procedures

- Consistent with the theory of action, the assessment administration process reflects non-linear and diverse ways that students learn and demonstrate their learning.
- Test administrators choose the content standards for assessment from the pool that meet a pre-specified set of criteria (e.g., “Choose three EEs from within Claim 1.”) to achieve blueprint coverage.
- For each selected content standard, testlet administration procedures use multiple sources of information to assign testlets, including student characteristics, prior performance, and educator judgment.




# Relevant Test Design and Development Statements in the Theory of Action

Three propositions are related to the depth, breadth, and complexity of administered assessments, as summarized in Table 10.11.

**Table 10.11:** *Propositions and Evidence for the Appropriate Combination of Testlets*

| Proposition | Procedural evidence | Empirical evidence | Type | Chapter(s) |
|---|---|---|---|---|
| First Contact survey correctly assigns students to appropriate complexity band | Description of First Contact survey design and algorithm development, First Contact helplet video | Pilot analyses, educator adjustment patterns | Content, Response Process | 4 |
| Administered testlets are at the appropriate linkage level | Description of testlet-level selection procedures, Instruction and Assessment Planner, administration fidelity,† mini-maps† | Educator selection and adjustment patterns, linkage level parameters and item statistics, educator focus groups | Content | 3, 4 |
| Administered testlets cover the full blueprint | Description of Essential Element selection procedures, Instruction and Assessment Planner, monitoring extracts, blueprints,† administration fidelity† | Blueprint coverage extracts and analyses, educator selection patterns, Special Circumstance codes | Content | 4 |

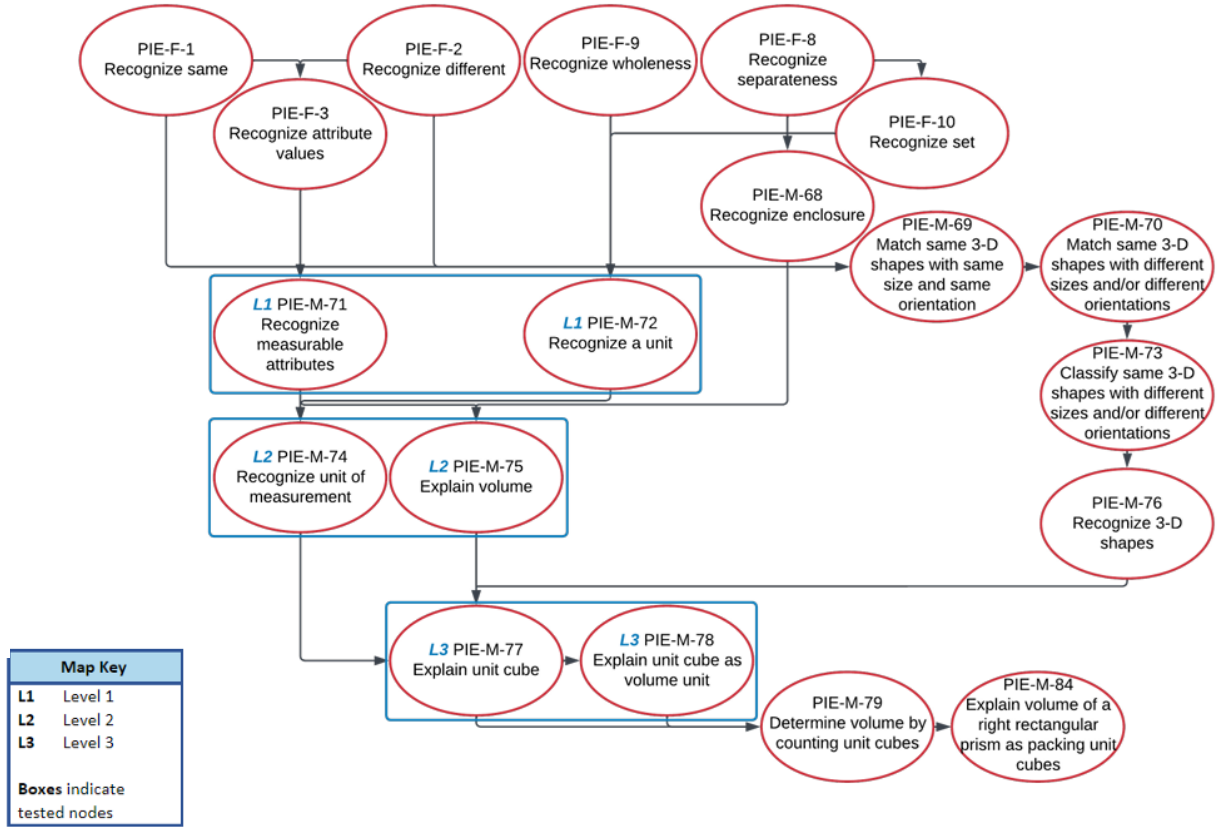
† Relies on evidence from Theory of Action input statements, as shown in Figure 10.8.

# CGSA - Pathways for Instructionally Embedded Assessment (PIE)

- CGSA funded grant project that began in fall 2022 led by the Missouri Department of Elementary and Secondary Education and in partnership with ATLAS.
- PIE is a four-year project aimed at designing, developing and evaluating a prototype integrated assessment model for 5th grade general education students in mathematics.

# CGSA - Pathways for Instructionally Embedded Assessment (PIE)

- Some features we are considering:
  - Assessments based on learning maps known as learning pathways
  - Teacher selection of standards to create content groupings as the basis of instruction and assessment
  - Pilot design includes full coverage of content standards in both instructionally embedded and end of year assessment administrations



# Responding to Peer Review Requirements



51

# IN GENERAL

- Don’t assume peers deeply understand your assessment design. Educate them.
  - Include in the submission index a succinct statement that “answers the question” the critical element is asking or provides. background needed to evaluate the evidence.
  - Leave “bread crumbs” in the index responses to cross-reference critical elements.
  - Explain atypical evidence.
  - Strive for coherence.

# Provide background needed to evaluate the evidence

**Notes**

**Validity Framework and Overall Evaluation**

The DLM validity framework is based in the project's theory of action (1.a.i), developed with state partners. There are four propositions to support the intended uses and interpretations of DLM scores:

1. Scores represent what students know and can do.
2. Achievement level descriptors provide useful information about student achievement.
3. Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.
4. Assessment scores provide useful information to guide instructional decisions.

Summative scores from DLM assessments are intended for use for several purposes (1.b.vi):

1. Reporting achievement and growth within the taught content aligned to grade-level content standards to a variety of audiences including educators and parents
2. Inclusion in state accountability models to evaluate school and district performance
3. Planning instructional priorities and program improvements for the following school year

Technical documentation of evidence supporting the validity of score interpretation and use includes material included in and referenced in chapters throughout the 2014-2015 Dynamic Learning Maps Technical Manual. The Manual addresses the design and development of the assessment, alignment of standards and test content, test administration, and test scores and reports. Evidence is presented related to content, response process, internal structure, relationships to other variables, and consequences (1.b.i). Evaluation of the evidence for overall validity of score interpretation and use is described for each proposition and related assumptions (1.b.ii), and is summarized in Chapter 11 of the Manual. Evaluation results indicate general support for the propositions and intended uses of summative results (1.b.iii), appropriate for the first year of a new assessment system. Additional validity studies are planned and in progress (1.b.iv) and additional procedural evidence is being collected as part of the consortium's continuous improvement process (1.b.v).

# Help peers interpret atypical evidence

Scoring is conducted at the linkage level within each Essential Element (EE), and an overall performance level is reported based on the total number of linkage levels mastered in the content area. Results are also reported for each EE and at the conceptual area level. Conceptual areas contain groups of related EE (content standards). Based on the diagnostic classification model, reliability evidence is provided at three levels:

- Content-area (performance-level) reliability provides reliability evidence for the total number of linkage levels mastered across all EEs for a given content area, which is analogous to total score reliability in Classical Test Theory (CTT)- or Item Response Theory (IRT)-based models. Estimates were calculated for each grade level in each content area, as demonstrated by the correlation between true and estimated number of linkage levels mastered. Values ranged from .909 to .965, indicating generally consistent measurement at the content area level (1.c.i).
- EE reliability provides reliability evidence for the number of linkage levels mastered within a single EE. Estimates were calculated for the 255 EEs across both content areas. EE reliability statistics are at a finer grain size than conceptual area because each conceptual area contains multiple EE. In this sense, conceptual areas are like strands and conceptual area results are like sub-scores. While conceptual area reliability estimates are planned for future analysis, EE reliability statistics provide evidence of consistency at the content standard level. Results from the Pearson correlation between true and observed values indicated that for 77.8% of EEs, the correlation was ≥ .75 (1cii).
- Linkage Level reliability provides reliability evidence for the classification accuracy of each of the 1,275 individual linkage levels across both content areas. Although at a larger grain size than item-level reliability statistics in CTT or IRT-based models, the linkage level is the smallest reported unit in a diagnostic classification model scoring system (1.a.i, 1.b.i). Results of the tetrachoric correlation between true and observed mastery status indicated that for 82.2% of linkage levels, the correlation was ≥ .80 (1.c.iii).

# Examples for Critical Element 2.1

- Briefly explain, provide evidence about test design immediately after purposes + uses, make sure the logic linking the two is clear.
  - Remind/point back to this response in 3.1.
- Ensure blueprint evidence is specific enough that peers can see breadth + depth of coverage, overall length supports score reporting and intended uses.
- CAT: item pool has enough breadth and depth to support the design.
- If applicable: Prove existing item banks can meet the need of the new system.

# Using innovative item types?

- 2.1: Explain how you know those items are appropriate to measure the knowledge & skills, depth and breadth of the standards
- 2.2: Don’t skimp on the procedural evidence that test development procedures successfully produce those items
- Cross-reference opportunities in 3.2, 4.2, 5.3

# In General (2)

- When the assessment system has multiple purposes and intended uses:
  - Evidence across the critical elements needs to address summative score uses.
  - Be clear about which parts of the system are subject to peer review when describing the evidence.

# Don’t have a theory of action? (CE 3.1)

- Talk early and often with TAC about how to plan, evaluate, and synthesize validity evidence.
  - First operational assessment year vs long-term plans.
  - Expected thresholds of evidence.
- Ensure consistent information across 2.1, 2.2 and 3.1 (content evidence).
- Cross-walk with 4.7 for areas with intended improvements.

# Q&A



59

# QUESTIONS?

# STILL MORE QUESTIONS?

- Submit your questions using the QR code.
- Attend session 1G (*Preparing for Assessment Peer Review*) Wednesday afternoon for answers.

# Thank You!