OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

SCHOOL SUPPORT AND ACCOUNTABILITY

2023 State Assessment Conference
1D - Overall Validity

September 26, 2023

FOCUS AREA: ASSESSMENT PEER REVIEW

1D - Overall Validity

Phoebe Winter, Independent Consultant, phoebe.winter@outlook.com
Brooke Nash, ATLAS, bnash@ku.edu
Chris Rozunick, Texas Education Agency, christine.rozunick@tea.texas.gov
Zach Warner, NYS Ed Dept, zachary.warner@nysed.gov
Nathan Dadey, Center for Assessment, ndadey@nciea.org

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 Introduction
- Argument Based Approaches to Validity
- Validity Arguments and Peer Review
- State and Consortium Examples
- Responding to Peer Review Requirements
- Q&A

Framing & Introduction

Overall Validity: Considerations for Multiple Approaches

The multiple approaches to assessment considered here typically have different or additional purposes and uses from traditional state summative assessments. To meet these purposes and uses, they may have between-student variation in

- What is assessed
- When students are assessed
- How they are assessed

Multiple Approaches Handout

Download for reference:

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

Which Critical Elements?

SECTION 3: TECHNICAL QUALITY - VALIDITY

- 3.1 - Overall validity, including validity based on content
  - The State’s academic assessments measure the knowledge and skills specified in the State’s academic content standards
- 3.2 - Validity based on cognitive processes
- 3.3 - Validity based on internal structure
- 3.4 - Validity based on relations to other variables

Professional Standards

Validity is the most fundamental consideration in developing tests and assessments (p. 11).

The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score and assessment results and interpretations (p. 11).

Source:
AERA, APA, & NCME (2014). Standards for educational and psychological testing. Washington, D.C.: AERA.

Focus of Session

For multiple approaches to assessment

- Validity arguments
- Evaluating claims with evidence
- Synthesizing evidence

Validity Reminders

- Validity is NOT a property of the test.
- Validity IS a property of the proposed interpretations of test scores for specific uses.
- Validity is a matter of degree.

Argument Based Approaches to Validity

Argument-Based Approach to Validity

Two basic steps to an argument-based approach to validity:

1. State the claims associated with the proposed interpretation or use → Interpretation and Use Argument
2. Evaluate the claims → Validity argument

Source:
Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73. doi: 10.1111/jedm.12000

Interpretation and Use Argument

- The logic model for making explicit the inferences, claims, and assumptions necessary to make links between the observed test score and intended interpretations and uses.
- Should be clear, coherent, plausible and comprehensive

Well articulated interpretation and use arguments include:

- Intended interpretations of results.
    - Claims or assumptions associated with the intended interpretations of results.
- Uses of results
    - Claims or assumptions associated with the intended uses of results.

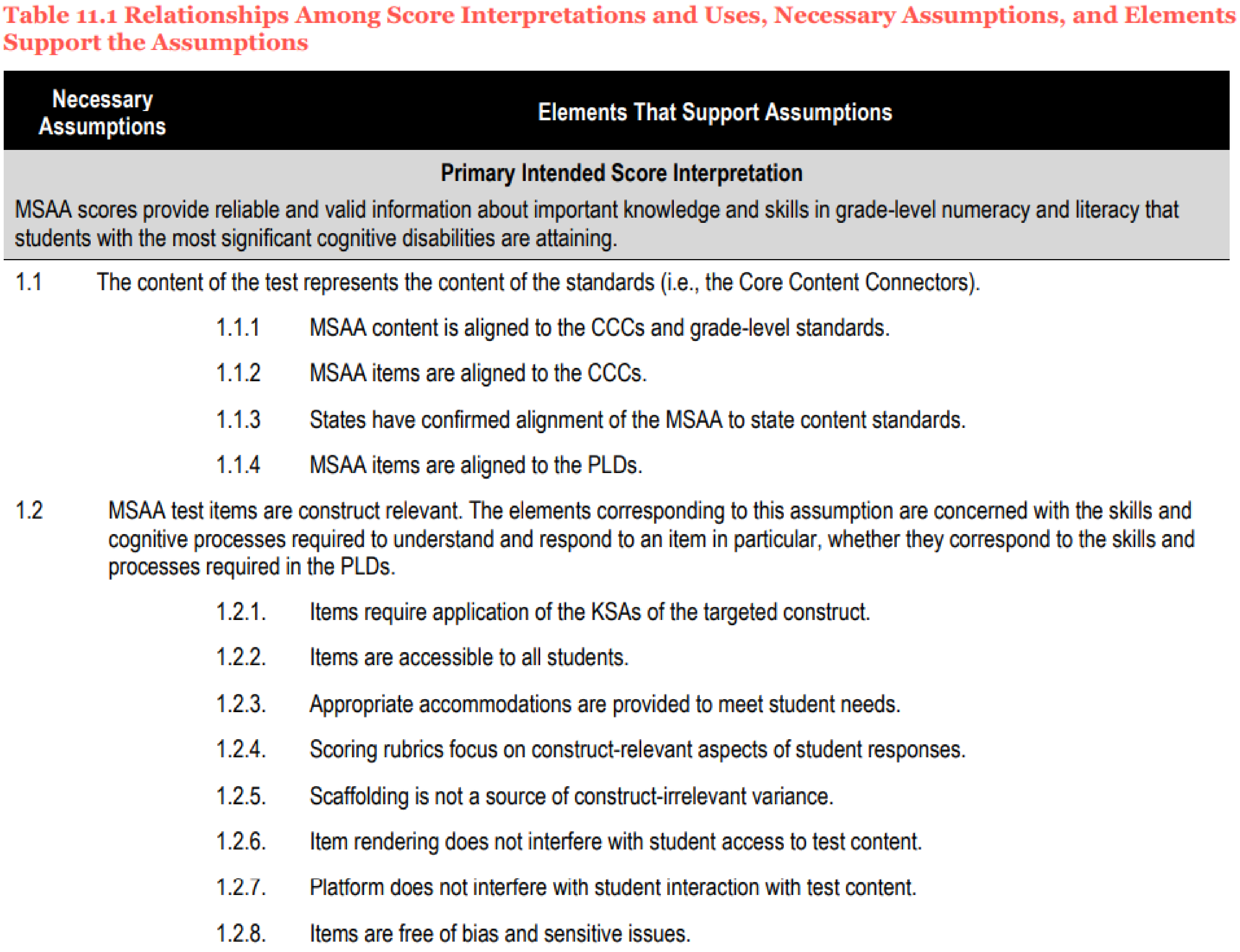


Table 11.1 Relationships Among Score Interpretations and Uses, Necessary Assumptions, and Elements That Support the Assumptions

Necessary Assumptions	Elements That Support Assumptions

Primary Intended Score Interpretation

MSAA scores provide reliable and valid information about important knowledge and skills in grade-level numeracy and literacy that students with the most significant cognitive disabilities are attaining.

1.1 The content of the test represents the content of the standards (i.e., the Core Content Connectors).

1.1.1 MSAA content is aligned to the CCCs and grade-level standards.

1.1.2 MSAA items are aligned to the CCCs.

1.1.3 States have confirmed alignment of the MSAA to state content standards.

1.1.4 MSAA items are aligned to the PLDs.

1.2 MSAA test items are construct relevant. The elements corresponding to this assumption are concerned with the skills and cognitive processes required to understand and respond to an item in particular, whether they correspond to the skills and processes required in the PLDs.

1.2.1. Items require application of the KSAs of the targeted construct.

1.2.2. Items are accessible to all students.

1.2.3. Appropriate accommodations are provided to meet student needs.

1.2.4. Scoring rubrics focus on construct-relevant aspects of student responses.

1.2.5. Scaffolding is not a source of construct-irrelevant variance.

1.2.6. Item rendering does not interfere with student access to test content.

1.2.7. Platform does not interfere with student interaction with test content.

1.2.8. Items are free of bias and sensitive issues.

Example Interpretation and Use Argument

Example taken from Multi-State Alternate Assessment - 2021 Technical Report

www.azed.gov/sites/default/files/2021/10/2020-21%20MSAA%20Technical%20Report_ADA.pdf

Multiple Approaches and Proposed Interpretations

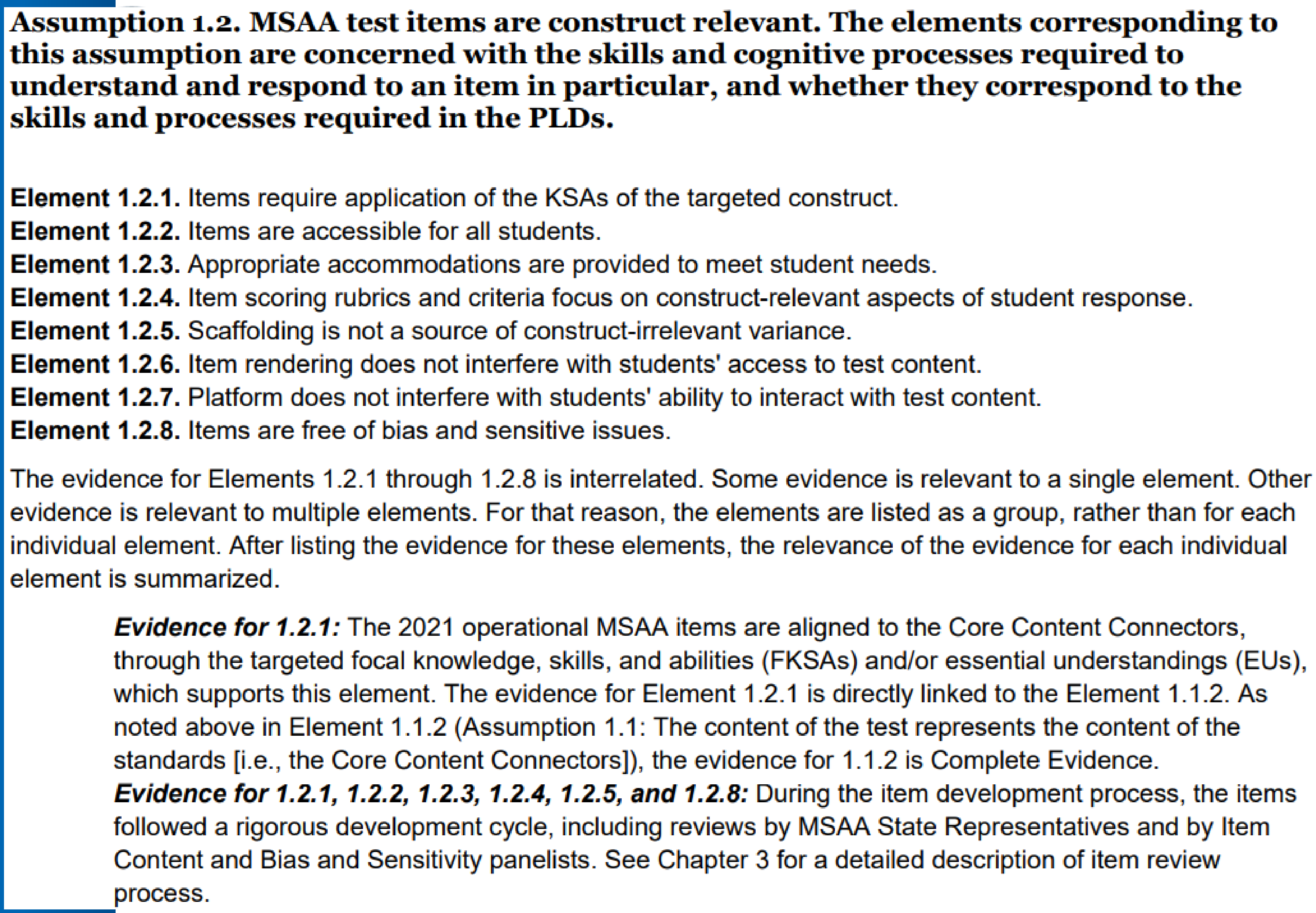
1. Test scores can have multiple possible interpretations/uses.
2. The validity of a proposed interpretation/use depends on how well the evidence supports the proposed interpretation/use.
3. More ambitious interpretations and uses requires more evidence.

Kane, 2013

Validity Arguments

- Provides an evaluation of all the claims and assumptions outlined in the interpretation and use argument.
- Well articulated validity arguments:
    - Situate collected evidence within each of the claims and assumptions in the interpretation and use argument.
    - Evaluate the degree to which the evidence supports each claim and assumption.
    - “Integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses” (p. 11 of the Standards).

Example Validity Argument



Assumption 1.2. MSAA test items are construct relevant. The elements corresponding to this assumption are concerned with the skills and cognitive processes required to understand and respond to an item in particular, and whether they correspond to the skills and processes required in the PLDs.

Element 1.2.1. Items require application of the KSAs of the targeted construct.
Element 1.2.2. Items are accessible for all students.
Element 1.2.3. Appropriate accommodations are provided to meet student needs.
Element 1.2.4. Item scoring rubrics and criteria focus on construct-relevant aspects of student response.
Element 1.2.5. Scaffolding is not a source of construct-irrelevant variance.
Element 1.2.6. Item rendering does not interfere with students' access to test content.
Element 1.2.7. Platform does not interfere with students' ability to interact with test content.
Element 1.2.8. Items are free of bias and sensitive issues.

The evidence for Elements 1.2.1 through 1.2.8 is interrelated. Some evidence is relevant to a single element. Other evidence is relevant to multiple elements. For that reason, the elements are listed as a group, rather than for each individual element. After listing the evidence for these elements, the relevance of the evidence for each individual element is summarized.

> ***Evidence for 1.2.1:*** The 2021 operational MSAA items are aligned to the Core Content Connectors, through the targeted focal knowledge, skills, and abilities (FKSAs) and/or essential understandings (EUs), which supports this element. The evidence for Element 1.2.1 is directly linked to the Element 1.1.2. As noted above in Element 1.1.2 (Assumption 1.1: The content of the test represents the content of the standards [i.e., the Core Content Connectors]), the evidence for 1.1.2 is Complete Evidence.
> ***Evidence for 1.2.1, 1.2.2, 1.2.3, 1.2.4, 1.2.5, and 1.2.8:*** During the item development process, the items followed a rigorous development cycle, including reviews by MSAA State Representatives and by Item Content and Bias and Sensitivity panelists. See Chapter 3 for a detailed description of item review process.

Example taken from Multi-State Alternate Assessment - 2021 Technical Report
www.azed.gov/sites/default/files/2021/10/2020-21%20MSAA%20Technical%20Report_ADA.pdf

Validity Argument and Theory of Action

- While theories of action are not necessary for developing an interpretation and use argument, they can be useful!
- Theories of action extend beyond interpretation and use claims to include claims about the intended change in stakeholder behavior.
    - A theory of action can provide a framework for development, use and evaluation of a new assessment program intended to elicit a positive change in learning and instruction (NCME, 2018).
- Other potential benefits:
    - Supports deep and robust articulations of use.
    - Helps explicitly define intended relationships between claims.

Validity Arguments and Peer Review

Validity Argument and Peer Review

Regardless of framework used to explicate and evaluate the validity argument, evidence of validity can be easily summarized according to the peer review Critical Elements (Ces) (if clearly articulated and comprehensive).

Establishing a strong validity argument supports all aspects of an assessment.

- Drives design decisions (including revisions).
- Supports item/task type selections.
- Connects content with results (e.g., claims/evidence, performance level descriptors (PLDs)).
- Guides reporting decisions.

State and Consortium Examples

NY Performance-based Assessments

NY also has a longstanding practice of including performance-based items in science examinations at all levels.

Purpose and Use

The science performance items allow students to demonstrate specific knowledge and skills articulated in the learning standards through hands-on laboratory experiences. These items make up 15% of the total test score.

The results of the NY Science Tests are intended to provide a measure of the extent to which individual students achieve the New York State Science Learning Standards for their course/grade level. In addition, the results are aggregated in order to determine whether schools, districts, and the State meet the required progress objectives specified in the New York State accountability system.

Although not currently planned to be used for accountability purposes under ESSA, NY's Performance-Based Learning and Assessment Networks (PLAN) Program is exploring the potential for New York's educational assessment strategy to be reimagined in a way that purposefully fosters high-quality instructional opportunities, provides authentic measures of deeper learning, and better prepares students for college and the workplace.

NY Performance-based Assessments

Theory of Action (ToA)

If students are presented with opportunities to demonstrate course/grade-level science knowledge and skills via hands on activities:

- their performance on these activities will produce evidence of their comprehension of the specific learning standards associated with those knowledge/skills and

- contribute to a total score that allows for inferences about student attainment of the learning standards for the course/grade level.

NY Performance-based Assessments

Types of Validity Evidence Needed for ToA

- Evidence that performance tasks are included and appropriate for the content (e.g., blueprints/content coverage, task/form specs – don't forget complexity info).

- Evidence that tasks are designed to solicit intended evidence (e.g., task development specs/processes, alignment studies).

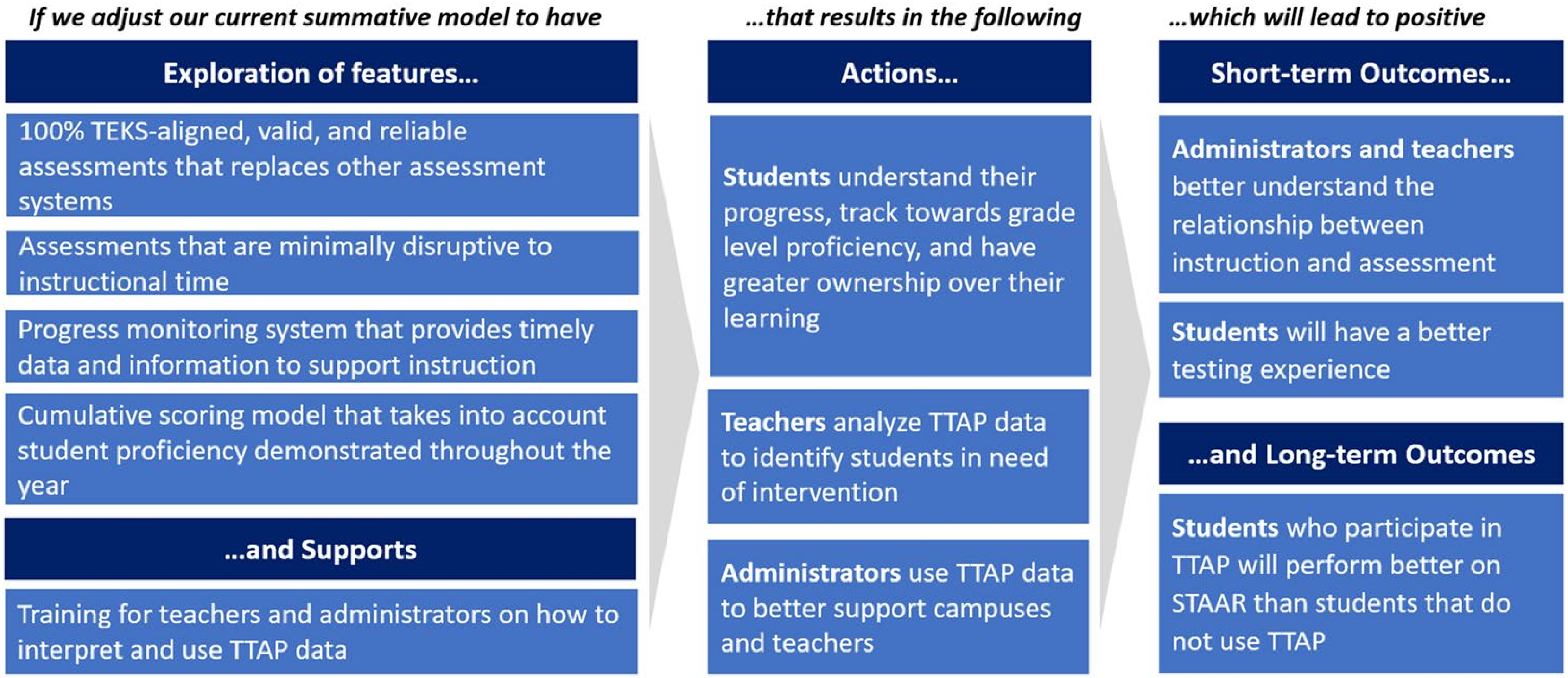
- Evidence that evidence produced by tasks informs the intended interpretations for all student groups (e.g., cog labs – different use from CE 3.2, technical reporting that connects results to ToA).

Texas Through Year Assessment Pilot

Pilot Overview:

- Multi-year pilot program to investigate the feasibility of Through Year Assessment in Texas.
- Currently working in Math, Science, and Social Studies and will add RLA a little later.
- In addition to the pilot, there is a very robust research agenda layered in to answer our many research questions.

Texas Through Year Assessment Pilot Theory of Action



If we adjust our current summative model to have

Exploration of features…

- 100% TEKS-aligned, valid, and reliable assessments that replaces other assessment systems
- Assessments that are minimally disruptive to instructional time
- Progress monitoring system that provides timely data and information to support instruction
- Cumulative scoring model that takes into account student proficiency demonstrated throughout the year

…and Supports

- Training for teachers and administrators on how to interpret and use TTAP data

…that results in the following

Actions…

- **Students** understand their progress, track towards grade level proficiency, and have greater ownership over their learning
- **Teachers** analyze TTAP data to identify students in need of intervention
- **Administrators** use TTAP data to better support campuses and teachers

…which will lead to positive

Short-term Outcomes…

- **Administrators and teachers** better understand the relationship between instruction and assessment
- **Students** will have a better testing experience

…and Long-term Outcomes

- **Students** who participate in TTAP will perform better on STAAR than students that do not use TTAP

Texas Through Year Assessment Pilot

Research Needed to Support the Theory of Action:

- Align performance levels to TTAP scale.
- Check percentile rank applications.
- Study reliability, validity, and comparability to STAAR to ensure that model could pass federal peer review standards (year 2+).
- Ensure routing performance of multi-stage test that reduces length of test while also providing reliable performance data.
- Score combinations for total score.
- Summative score comparisons to TTAP pilot results - checking across student demographics.
- Comparison of student performance between TTAP and non-TTAP users (matched study).

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 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

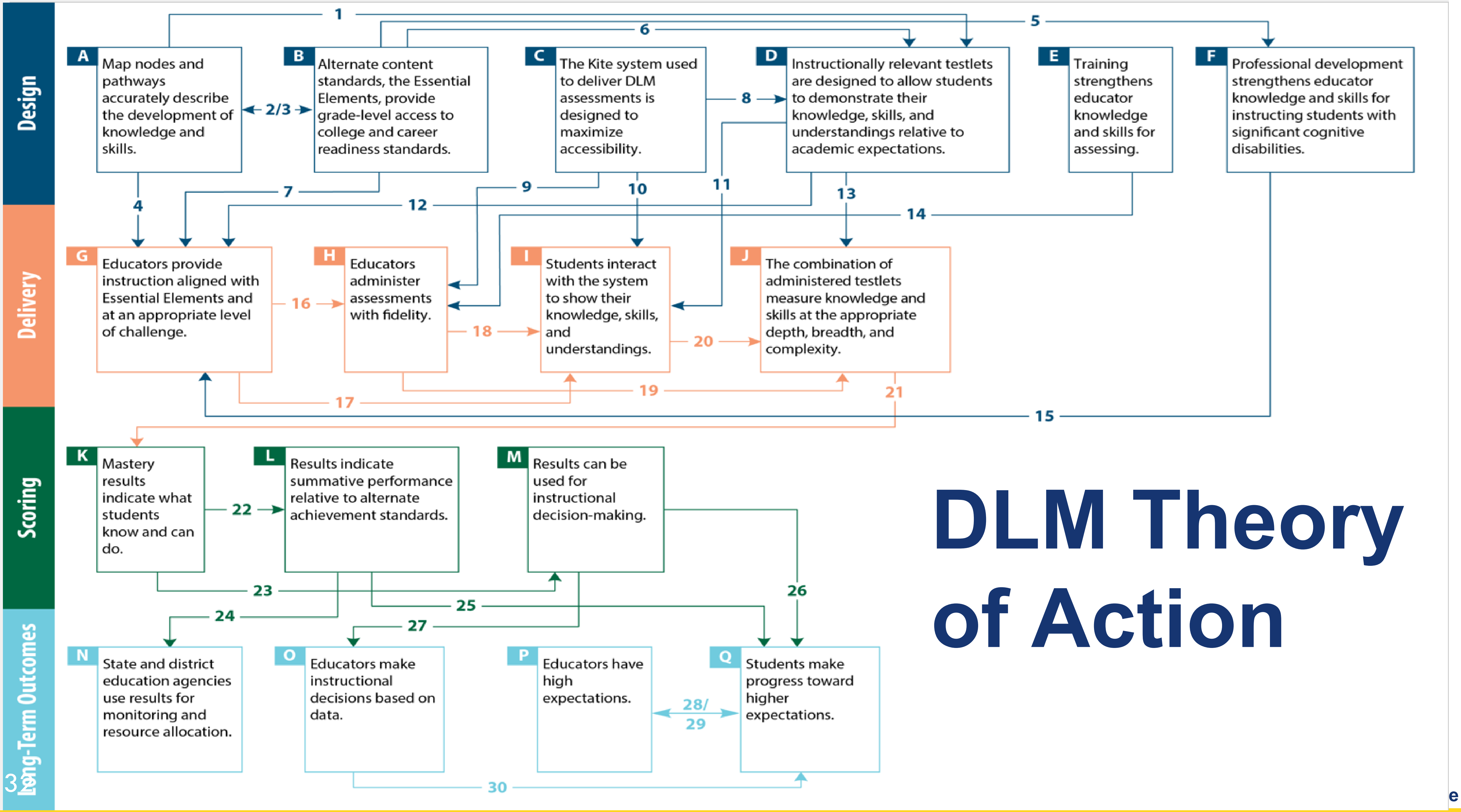
- Based on learning maps that describe how students acquire knowledge and skills.
  - The learning maps provide a framework that supports inferences about student learning needs.
- A set of learning targets for instruction and assessment aligned to grade-level academic content standards.
- Instructionally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

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

DLM started with a theory of action…

- The DLM theory of action was formulated in 2011, revised in 2013 and revised again in 2019.
- It rests on the belief that high expectations for students with the most significant cognitive disabilities, combined with appropriate educational supports and diagnostic tools for educators, results in improved academic experiences and outcomes for students and educators.
    - The theory of action was used to define how the system was going to elicit important changes for students!

DLM Theory of Action



A Map nodes and pathways accurately describe the development of knowledge and skills.

B Alternate content standards, the Essential Elements, provide grade-level access to college and career readiness standards.

C The Kite system used to deliver DLM assessments is designed to maximize accessibility.

D Instructionally relevant testlets are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations.

E Training strengthens educator knowledge and skills for assessing.

F Professional development strengthens educator knowledge and skills for instructing students with significant cognitive disabilities.

G Educators provide instruction aligned with Essential Elements and at an appropriate level of challenge.

H Educators administer assessments with fidelity.

I Students interact with the system to show their knowledge, skills, and understandings.

J The combination of administered testlets measure knowledge and skills at the appropriate depth, breadth, and complexity.

K Mastery results indicate what students know and can do.

L Results indicate summative performance relative to alternate achievement standards.

M Results can be used for instructional decision-making.

N State and district education agencies use results for monitoring and resource allocation.

O Educators make instructional decisions based on data.

P Educators have high expectations.

Q Students make progress toward higher expectations.

e

DLM Validity Argument

Three-tiered approach to assessment validation:

1. The theory of action defines the statements or claims that must be in place to achieve the goals of the system (which encompass the intended uses).
2. The interpretive argument defines the propositions that must be evaluated to support each statement or claim in the theory of action.
3. Validity studies are identified to evaluate each proposition in the interpretive argument.

https://2022-ie-techmanual.dynamiclearningmaps.org/10-validity-argument

Summarizing Validity Evidence

- Evidence is summarized for each statement in the theory of action and for each proposition underlying the statement.
- Evidence is further categorized according to the five types of evidence for validity arguments defined by the Standards (content, response process, internal consistency, relation to other variables, and consequences.

Example Evidence Summary

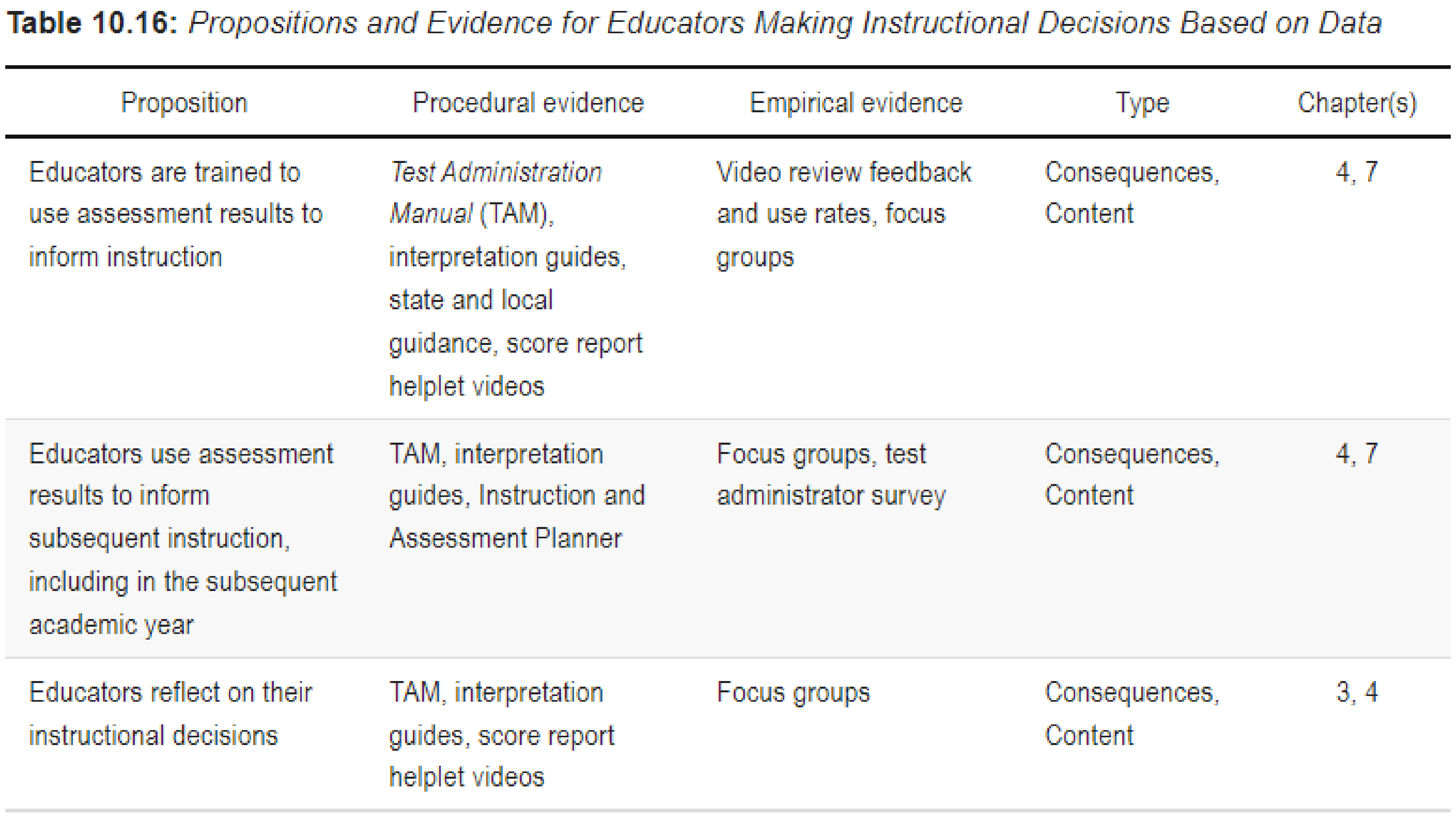


Table 10.16: *Propositions and Evidence for Educators Making Instructional Decisions Based on Data*

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Educators are trained to use assessment results to inform instruction	*Test Administration Manual* (TAM), interpretation guides, state and local guidance, score report helplet videos	Video review feedback and use rates, focus groups	Consequences, Content	4, 7
Educators use assessment results to inform subsequent instruction, including in the subsequent academic year	TAM, interpretation guides, Instruction and Assessment Planner	Focus groups, test administrator survey	Consequences, Content	4, 7
Educators reflect on their instructional decisions	TAM, interpretation guides, score report helplet videos	Focus groups	Consequences, Content	3, 4

For more information, please visit the DLM website

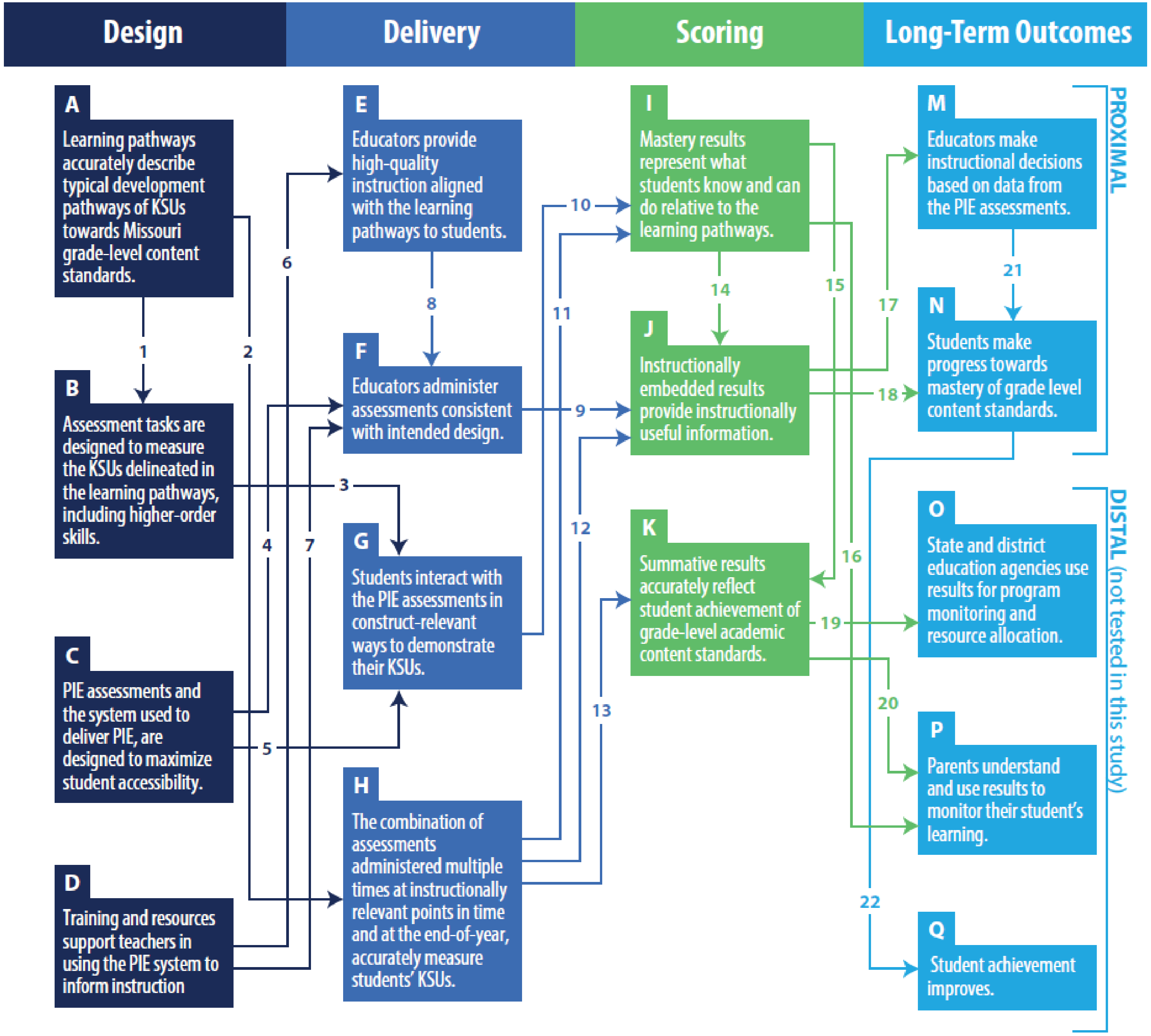
CGSA - Pathways for Instructionally Embedded Assessment (PIE)

- CGSA funded grant project that began in fall 2022
- PIE is a four-year project aimed at designing, developing and evaluating a prototype integrated assessment model for 5th grade mathematics.
- The project's ultimate goal is to use data collected from the integrated model (instructionally embedded + end of year assessments) to evaluate its use, or a variation of it, for future potential use as a statewide summative assessment model.

PIE Project Goals

1. Design, develop, administer, and evaluate the PIE Assessment System based on learning pathways aligned to grade-level content standards.
2. Evaluate the usability of the PIE Assessment System under natural conditions.
3. Design an approach to evaluating technical adequacy, including scoring model, theory of action, and validation plan for future use as a statewide assessment.
4. Broadly disseminate project materials and findings to a variety of audiences, including the proof of concept for future use as a statewide assessment system.

Draft PIE Theory of Action



Design

A Learning pathways accurately describe typical development pathways of KSUs towards Missouri grade-level content standards.

B Assessment tasks are designed to measure the KSUs delineated in the learning pathways, including higher-order skills.

C PIE assessments and the system used to deliver PIE, are designed to maximize student accessibility.

D Training and resources support teachers in using the PIE system to inform instruction

Delivery

E Educators provide high-quality instruction aligned with the learning pathways to students.

F Educators administer assessments consistent with intended design.

G Students interact with the PIE assessments in construct-relevant ways to demonstrate their KSUs.

H The combination of assessments administered multiple times at instructionally relevant points in time and at the end-of-year, accurately measure students' KSUs.

Scoring

I Mastery results represent what students know and can do relative to the learning pathways.

J Instructionally embedded results provide instructionally useful information.

K Summative results accurately reflect student achievement of grade-level academic content standards.

Long-Term Outcomes

M Educators make instructional decisions based on data from the PIE assessments.

N Students make progress towards mastery of grade level content standards.

O State and district education agencies use results for program monitoring and resource allocation.

P Parents understand and use results to monitor their student's learning.

Q Student achievement improves.

*KSUs = knowledge, skills and understandings

Responding to Peer Review Requirements

Explain Your Approach

- Purpose and use
  - Clearly describe purpose and use.
  - Explicitly tie evidence for relevant CEs to purpose and use.
- Using scores
  - Define how scores are used to meet peer review requirements.
  - Tie level of comparability to how scores are used for peer review purposes.
- Uses outside of peer review purview
  - If any non-peer reviewed purpose or use affects how a CE is addressed, explain as needed.
  - Clearly delineate non-peer reviewed purposes and uses if they are discussed.

Goal: Peer reviewers who understand the system and understand the reasons for using evidence that may be atypical.

Connect Evidence to Purpose and Use

- CE 3.1 Content, Content Balance, Cognitive Complexity, Depth and Breadth
- CE 3.2 Cognitive Processes
- CE 3.3 Internal Structure
- CE 3.4 Relationship to Other Variables

QUESTIONS?

STILL MORE QUESTIONS?

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