OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

SCHOOL SUPPORT AND ACCOUNTABILITY

2023 State Assessment Conference

1B: Alignment

September 26, 2023

FOCUS AREA: ASSESSMENT PEER REVIEW

1B - Alignment

Meagan Karvonen, ATLAS, karvonen@ku.edu
Phoebe Winter, Independent Consultant, phoebe.winter@outlook.com
Brooke Nash, ATLAS, bnash@ku.edu
Zach Warner, NYS Ed Dept, zachary.warner@nysed.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
- Responding to peer review requirements
- Q&A

Where is alignment in peer review?

Primary critical elements (CEs):

2.1: Test design and development

3.1: Overall validity, validity based on content

Also here:

2.2: Item development

4.7: Technical analysis, ongoing maintenance

6.3: Aligned academic achievement standards

(Not addressed in this session)

Alignment may also impact other CEs

Examples:

- 3.3: Internal structure
- 3.4: Relationship to other variables
- 4.3: Reliability (continuum)
- 4.5: Multiple forms

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.

3.1: Overall Validity + Content Evidence

Assessments measure the knowledge and skills in the content standards, including:

- Documentation of adequate alignment between the assessments and the content standards in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity;
- Documentation that the assessments address the depth and breadth of the content standards.

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.

4.7: Technical Analysis + Maintenance

- The state has a system for monitoring, maintaining, and improving, as needed, the quality of its assessment system, including clear and technically sound criteria for the analyses of all of the assessments in its assessment system.

Minimum Alignment Expectation

- Assessment measures depth and breadth of the standards.
- Balance of content (knowledge, cognitive process, cognitive complexity).

Common Misconceptions

- Alignment is only about the relationship of items to content standards.
- Alignment evidence comes from an external study on an operational item or task pool.
- Peers only accept alignment evidence that uses Webb's methods and criteria.

Example Challenge

Evidence could show poor alignment when Webb (1997) criteria assume:

- Intended equal coverage of content standards.
- All items count equally.
- Test length isn't a consideration (6 items per standard).
- There is nothing in the chain between standards and items.

In This Session

Issues and options for three types of designs and some specific examples:

- Through-year
- Portfolio, project-based, performance
- Matrix sampling

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).

Portfolio, project based or performance assessment*

- Application of knowledge, skills and abilities to **authentic problems**
- Require the student to **produce something** (e.g., a report, product, experiment, or performance)
- Scored against specific criteria
- May take place over **varying time periods**: of hours, days or weeks depending on the range and complexity of skills to be assessed

*Under ESEA, State assessments "... **may be partially delivered in the form of** portfolios, projects, or extended performance tasks" ESSA, Sec 1111(b)(2)(B)(vi)

Matrix Sampling of Assessed Content

- Each student receives a subset of the item content and associated standards through within the same administration window
  - Statistical sampling of items, subject to constraints, across students within the same year
  - E.g., sampling items or standards within content domains
  - Students receive overall achievement scores, based on a representative sample of the content
- Reduces testing time and information at the student-level, while still maintaining information at the school-level

Multiple Approaches Handout

Download for reference:

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

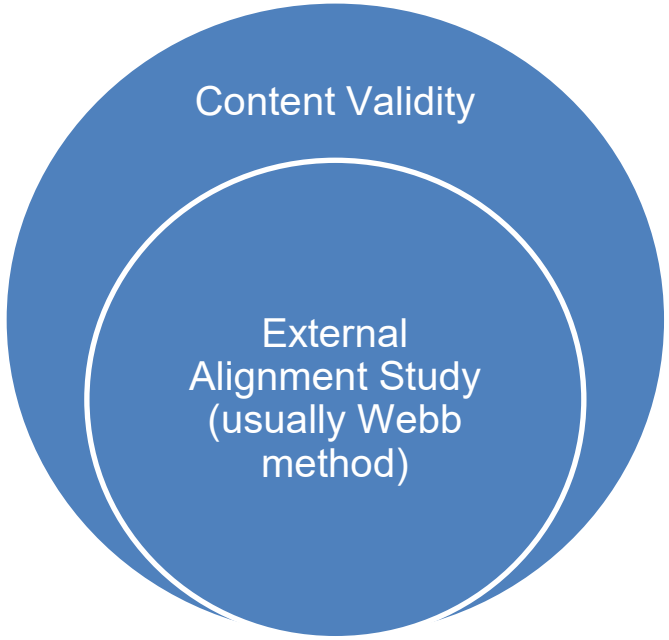
Alignment Possibilities

Multiple approaches to assessment may bring different ways of defining and evaluating alignment.

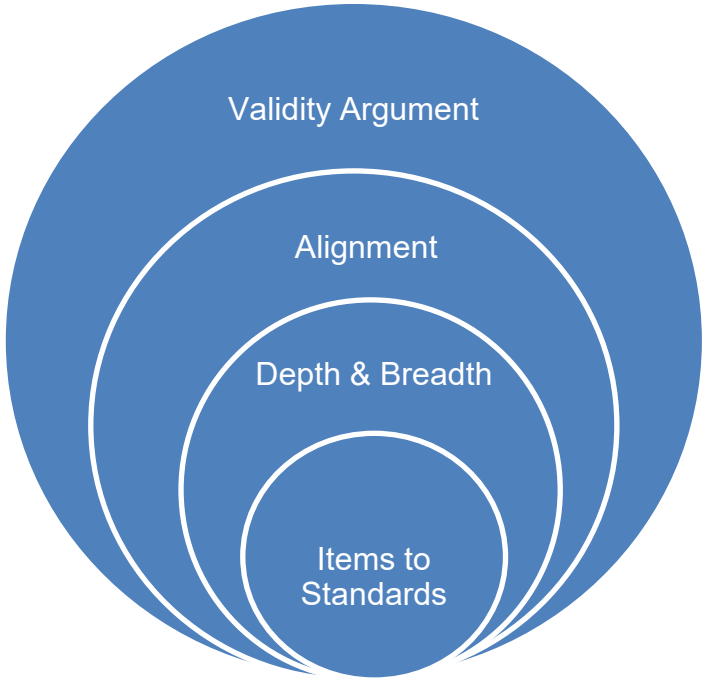
- Different components to align.
- Different thresholds of “adequate” alignment.
- Different view of alignment in relation to validity.

Issues and Options

ALIGNMENT AND VALIDITY - TRADITIONAL VIEW



ALIGNMENT AND VALIDITY - COMPREHENSIVE VIEW



Five Sources of Validity Evidence:

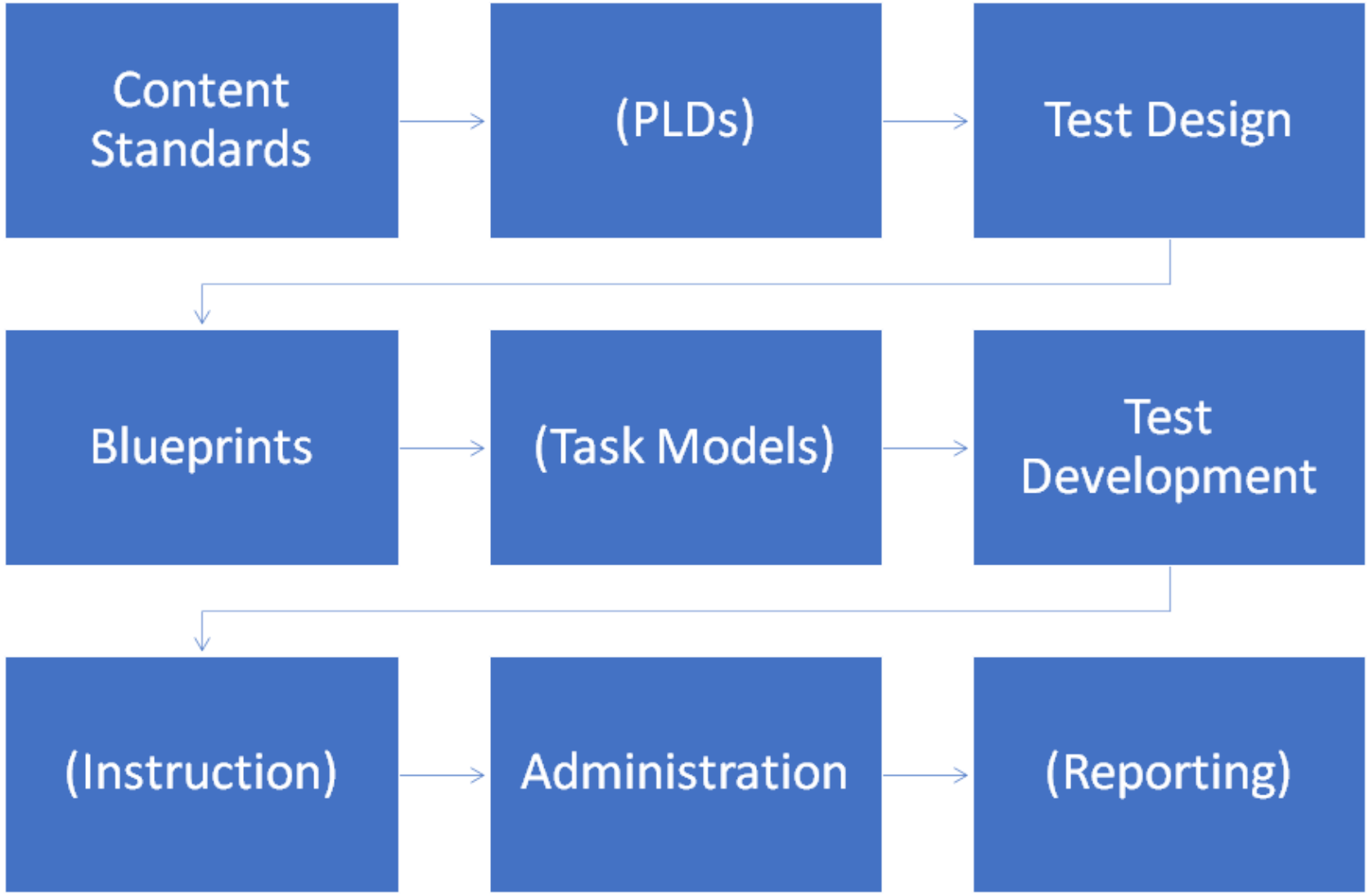
1. Content
2. Cognitive Process
3. Internal Structure
4. Relationships with Other Variables
5. Consequences

*Claims about alignment within a coherent system requires evidence within multiple sources of validity evidence.

BENEFITS OF COMPREHENSIVE ALIGNMENT APPROACH

1. Supports development and implementation of well-articulated validity arguments and validation plans (which makes accumulating and synthesizing validity evidence for peer review much easier).
2. Accumulating alignment evidence throughout the design and development process can provide opportunities for identifying potential alignment issues earlier on in the development process (i.e., in time to correct prior to operational administration).
3. Establishing alignment expectations as part of the test design, supports external partners in designing and conducting alignment studies that are consistent with design.
    a. Also makes reconciling and addressing alignment findings easier!
4. Ultimately, more directly supports score interpretations and uses than the traditional view!

Potential Alignment Relationships



Through-Year/Instructionally Embedded Models (1)

Potential design features to consider when developing alignment plans:

- Plans for scoring and reporting after each administration (embedded or interim) and end of year (summative)
  - Are the planned reporting metrics aligned to intended uses and interpretations?
- Blueprint specification within and across assessment windows.
  - Are the specifications going to produce forms that are aligned to scoring and reporting plans?

Through-Year/Instructionally Embedded Models (2)

Potential design features to consider when developing alignment plans (cont.):

- Size and scope of item bank.
  - What are the plans for evaluating and collecting evidence of item alignment in the most efficient but comprehensive way?
  - Will content be available at multiple levels of complexity? If so, what are the expectations for alignment?
- Year-round administration of short tests/testlets.
  - How will alignment to blueprint standards across the totality of assessments administered by demonstrated?

Through-Year (TY)/Instructionally Embedded (IE) Models (3)

- Example types of evidence that could be used to demonstrate alignment of assessments with academic content standards for TY/IE models.
  - Organized by procedural and evaluative evidence types.

Through-Year/Instructionally Embedded Models (4)

Procedural:

- Description of content structure and coverage of content, as specified by test blueprints:
  - For example, one “master blueprint” that combines coverage across all assessment windows OR window-specific blueprints that articulate coverage in each (depends on intended use of results in summative calculations) OR both.
  - Also, including any flexibility users have in content selection and the impact that flexibility has, if any, on the inferences that can be made from results.

Through-Year/Instructionally Embedded Models (5)

Procedural (cont.):

- Description of ALDs/PLDs development process and procedures (including clear articulation of student performance expectations achieved **by** the end of the year).
- Description of score report (or reporting dashboard) design and the scoring models and procedures consistent with intended reporting metrics.

Through-Year/Instructionally Embedded Models (6)

Procedural (cont.):

- Evidence-centered design method to develop high-quality items aligned to targeted content.
- Item writer training on evidence-centered design and alignment.
- Formative alignment checks prior item field testing.
- Description of procedures for monitoring of blueprint coverage (i.e., particularly within the level of administration flexibility allowed by the approach).

Through-Year/Instructionally Embedded Models (7)

Evaluative:

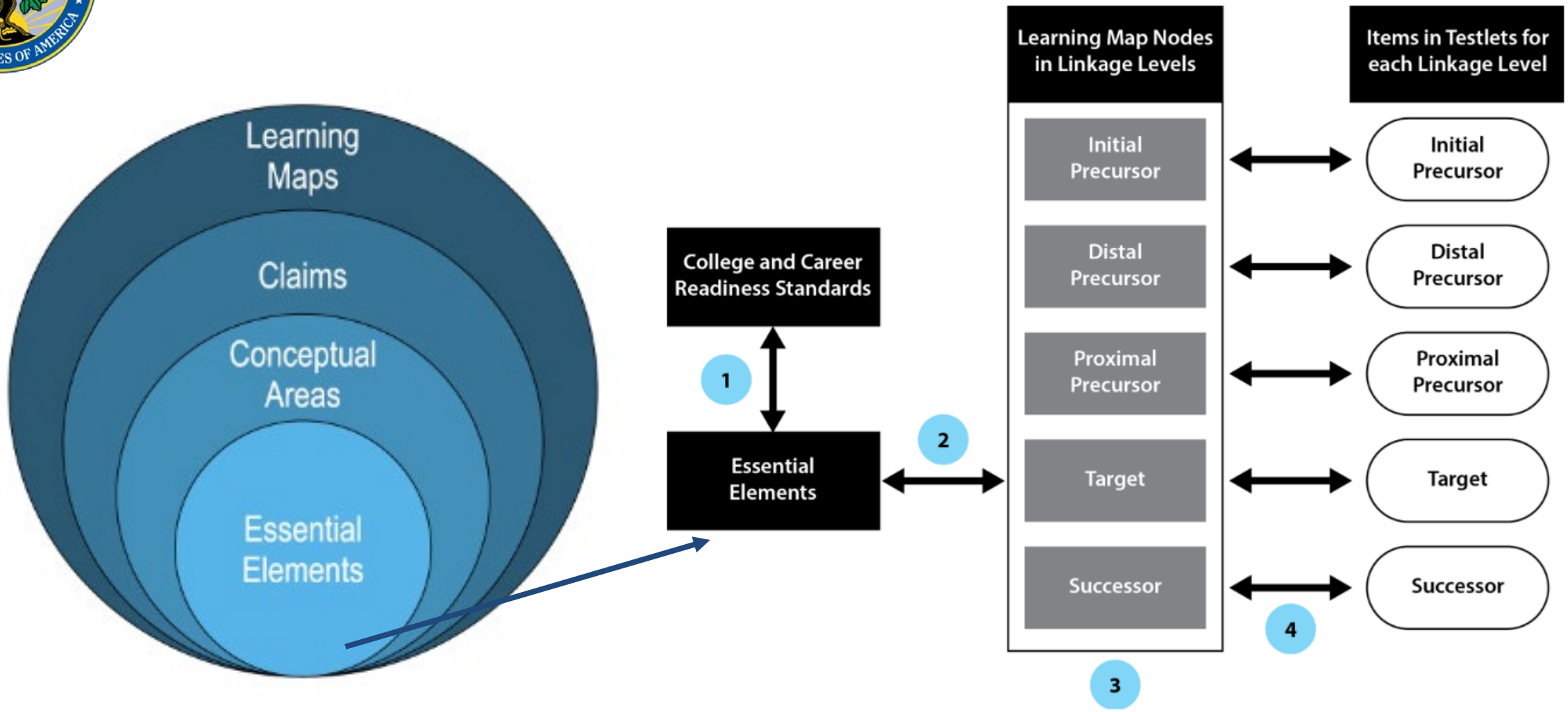
- Results from educator review of items (i.e., that occurred prior to field testing and were selected for operational forms).
- Analysis of blueprint coverage (i.e., particularly within the level of administration flexibility allowed by the approach).

Through-Year/Instructionally Embedded Models (8)

Evaluative (cont.):

- External alignment study to evaluate relationships between the content structures consistent with the design of the assessment. For example:
  - Progression of levels of complexity and alignment of items intended to measure those levels.
  - Sample of items from the pool to evaluate their alignment to targeted content.

DLM Content Structures



Considerations: Performance Assessment

Design features of performance based assessments to consider when developing an alignment study:

- Cover depth, but not necessarily breadth of content
- Variation within tasks (what aspects are consistent for all students, what is flexible)?
- Complexity is woven throughout the task and may need to be described differently.
- Administration time can vary widely.
- Scoring criteria (these are often the clearest articulation of what is info is valued from the task).

Considerations: Performance Assessment

Procedural:

- Domain analysis or “unpacking” of learning standards to show what knowledge/skills are inherent in the standards that are best assessed by a performance task (as opposed to selected-response or other item).
- Task specifications/blueprints, including any flexibility users have in content selection and the impact that flexibility has, if any, on the inferences that can be made from results.
- Information related to task development including training for writers and proctors.

Considerations: Performance Assessment

Evaluative:

- Results from educator review of tasks reporting their judgment on the content and cognitive process students will use in responding.
- External alignment study evaluating the relationships between the tasks and content structures (including not only the learning standards, but also the domain analysis, PLDs, or other "unpackings" of content).
- Demonstration that flexibility within administration does not hinder the measurement of the intended content/processes.

Considerations: Matrix Sampling

- How will breadth and depth of learning standards be covered 1) within each form/year and 2) across forms/years to ensure full coverage?

- Will the coverage support the necessary/desired level of reporting?

Matrix Sampling

Potential peer questions/concerns

- Why is matrix sampling appropriate for the assessment program?
- Can the blueprint support everything that comes downstream?
- (If also adaptive): how can we be sure each student receives an aligned test?

Matrix Sampling

Procedural evidence:

- Rationale for the blueprint.
- Evidence that combined blueprints cover the breadth and depth of the learning standards.
- Evidence that students take an assessment with sufficient breadth of coverage each year.
- Description of test development steps that promote alignment (task templates, item writer training, extended response item ratings, etc.).

Suggestions for Responding to Peer Review Requirements

Suggestions for Peer Review

1. Describe intended content relationships.
2. Describe procedural evidence.
3. Conduct an external alignment study using a design and criteria appropriate for the assessment.
4. Provide evidence of how the state interprets and responds to findings.

Intended Relationships

During test design process, describe intended content relationships.

- What parts are aligned?
- What should alignment look like (e.g., intended uneven coverage of standards)?
- How does the intended content coverage support reporting and score uses?

Procedural Evidence

Gather, summarize, and interpret procedural evidence. Examples:

- Stakeholder involvement, expert feedback on decisions.
- Rationales for design of elements in the system (how they promote alignment).
- Qualifications and training of test developers.
- Quantifiable procedural evidence (e.g., ER panel alignment ratings, CAT simulation studies).

External Alignment Study

Conduct an external alignment study using a design and criteria appropriate for the assessment.

- Ask the right alignment evaluation questions.
- Modify existing methodologies as needed.
- Identify a priori hypotheses, definitions of adequate (expected) alignment.
- Interpret findings in relation to expected alignment.

External Alignment Studies

- What content relationships need to be evaluated, in what ways?
  - *Hopefully not new and different from development phase*
- What cognitive taxonomy is used to describe depth of knowledge? What constitutes “higher-order” thinking skills?
- CAT examples
  - Sampling item pool to provide population estimates of overall alignment.
  - CAT algorithm delivers an aligned assessment for each student.

External Studies: A Range of Methods

Common methods:

- Webb
- Achieve
- Surveys of Enacted Curriculum
- Links for Academic Learning
- Generalized Assessment Alignment Tool
- Tailored Designs

…And Some Guidance

Examples:

- Evaluating alignment in large-scale standards-based assessment systems (Forte, 2017).
- A proposed framework for evaluating alignment studies (Davis-Becker & Buckendahl, 2013).

…And Also Methodological Options

Examples:

- The Relationship between Item Developer Alignment of Items to Range Achievement-Level Descriptors and Item Difficulty: Implications for Validating Intended Score Interpretations (Schneider et al., 2022).
- Gauging Uncertainty in Test-to-Curriculum Alignment Indices (Traynor et al., 2020).
- Evaluating Content-Related Validity Evidence Using a Text-Based Machine Learning Procedure (Anderson et al., 2020).

Responding to External Study

Provide evidence of how the state interprets and responds to findings (CE 4.7), such as:

- Next steps.
  - More evaluation where results need explanation.
  - Changes to future test development procedures.
  - Plan to write new items, re-evaluate alignment.
- TAC recommendations.

General Recommendations

- Explain the relevant content relationships each time you present the evidence.
- Support qualitative statements with quantitative data.
- Synthesize the evidence (procedural, empirical, from all stages) to make the case for how alignment goals are met, relative to peer review requirements (and additional validity claims if relevant).

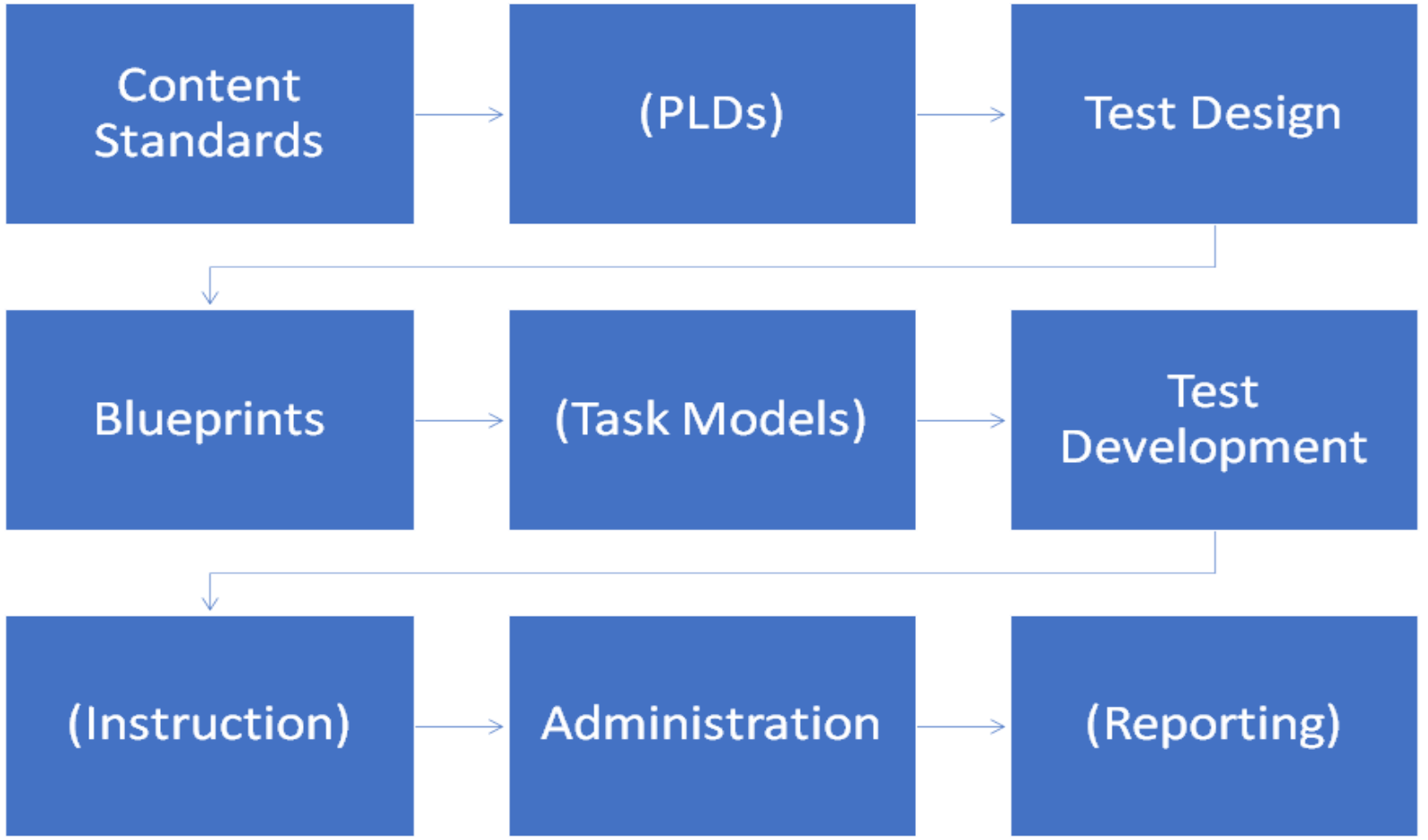
Let's apply these ideas

1. Access the **alignment activity handout**
2. Pick one of the assessment programs highlighted in the **multiple approaches handout** (or another of your choosing)
3. Alone or in small groups, discuss:
   - Which alignment questions are potentially appropriate for that assessment model?
   - What other information would you need to know before going further with a particular alignment approach?

Alignment Activity Handout

Multiple Approaches Handout

Remember the Potential Alignment Relationships



Questions to Consider

- Which alignment questions are potentially appropriate for your chosen assessment model?
- What other information would you need to know before going further with a particular alignment approach?

Alignment Activity

Multiple Approaches

Q&A

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!