OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

SCHOOL SUPPORT AND ACCOUNTABILITY

2023 State Assessment Conference

1F - Achievement Level Descriptors and Score Reporting

September 26, 2023

FOCUS AREA: ASSESSMENT PEER REVIEW

1F - Achievement Level Descriptors and Score Reporting

Nathan Dadey, Center for Assessment, ndadey@nciea.org
Meagan Karvonen, ATLAS, Karvonen@ku.edu
Zach Warner, NYS Ed Dept, zachary.warner@nysed.gov
Chris Rozunick, Texas Educational Agency, Christine.Rozunick@tea.texas.gov

Outline

1. Framing and overview (15 min)
2. Issues and options (30 min)
3. Responding to peer review requirements (15 min)
4. Q&A (15 min)

1. Framing & Overview

Which Critical Elements?

6.1 – Adoption of Academic Achievement Standards for All Students

6.2 - Achievement Level Standards Setting

6.3 - Challenging and Aligned Academic Achievement Standards

6.4 - Reporting

In Summary

These elements are focused on defining what academic achievement on the standards is, and then communicating that achievement to the field.

Defining Standards (6.1)

Develop State Academic Standards, Adopt Standards, Define Performance & Assessable Standards

P/ALD Development (6.2, 6.3)

Define the P/ALDs, in part by determining what aspects of student performance should define varying levels of performance

Standards Setting (6.2)

Translate the ALDs into cuts and corresponding levels through a standards-setting method

"Program" Reporting & Interpretive Materials (6.4)

Direct to the field reports, including individual level student reports (ISRs), and reports at other levels (including interpretive materials and logistics

State Reporting & Interpretive Materials (6.4)

Reporting through state level data systems, often in the terms of state report cards or data dashboards

All of these are specific "slices" of the assessment design, implementation and reporting process.

With connections to and implications for:

- 2.1 - Test Design and Development
- 2.2 - Item Development
- 3.1 - Overall Validity
- 3.2 - Validity Based on Cognitive Processes
- 3.3 - Validity Based on Internal Structure

Crash Course on Terminology

- **Academic Content Standards**: define what students should know and be able to do in a given domain or discipline
- **Performance or Achievement Level Descriptors (P/ALDs)**: Summary statements, typically across standards, of what performance or achievement is at multiple levels of performance
  - **Policy**: high level, communicates what it means to be, for example, college and career ready. Typically consistent across grades, excluding high school, and subject areas
  - **More Detailed ALDs:** Additional ALDs are often developed that articulate in content specific ways, knowledge, skills and abilities within a given level (e.g,. Range, Threshold/Target and Reporting ALDS).
- **Achievement Standards**: The cutscores created through a standards setting method and used to categorize students into the achievement levels.

Deeper Dive: 6.3 - Challenging and Aligned Academic Achievement Standards

The State's academic achievement standards are challenging and aligned with the State's academic content standards … [and the knowledge and skills necessary for success in college and the workforce]

1. **Challenging Achievement Level Standards.** The performance or achievement level descriptors (P/ALDs) present challenging knowledge and skills, the achievement standards (cut scores) are challenging, and AA-AAAS ALDs represent the highest levels of appropriate achievement.
2. **Meaningful Achievement Level Standards**. The P/ALDs are vertically articulated, differentiate between levels of achievement, "proficient" and above P/ALDs in high school represent the knowledge and skills necessary for success in college and the workforce
3. **Aligned Achievement Standards.** The P/ALDs represent are clearly derived from and represent the full range of the state content standard and the standards setting process considered the full range of state content standards

Deeper Dive: 6.3 - Challenging and Aligned Academic Achievement Standards

Common Sources of Evidence:

1. **Challenging Achievement Level Standards.** The ALDs, a summary or report of the ALD development process, expert opinion for AA-AAAS
2. **Meaningful Achievement Level Standards.** Documentation of vertical articulation processes for P/ALDs and standards setting, summaries of students in each performance level, item mapping studies, comparisons to external benchmarks, TAC minutes
3. **Aligned Achievement Standards.** Documentation of the ALD development process, crosswalk between ALDs and content standards, alignment analyses

Common Pitfalls:

- Challenges with deriving achievement level labels that are meaningful, support intended interpretations, and all stakeholders can agree on.

Deeper Dive: 6.2 - Achievement Standards Setting

The State used a technically sound method and process that involved panelists with appropriate experience and expertise for setting.

1. **Standards Setting.** The state used a technically sound and well-documented process to develop reasonable, defensible P/ALDs and achievement standards (i.e., cut scores), the cut scores are adequately reliable, and the cut scores distinguish between the levels described by the P/ALDs.
2. **Standard Setting Participants**. Participants had appropriate expertise, participants were representative, and participants were able to apply their knowledge and skill

Deeper Dive: 6.2 - Achievement Standards Setting

Common Sources of Evidence:

1. **Standards Setting**. P/ALD development process/script, standard setting process/script, standard setting training materials/script, standard setting report, TAC minutes, participant rating summaries
2. **Standard Setting Participants.** Summary of P/ALD development and standard setting participant characteristics, participant selection criteria

Deeper Dive: 6.2 - Achievement Standards Setting

Common Pitfalls:

- Participants who, collectively, do not represent those with expertise and understanding of all students (e.g., no teachers with experience with EL or SpEd students)
- Not providing enough detail or justification for the standards setting process
- Not allowing enough time for the standards setting process or clearly explaining the process during training
  - Not using the P/ALDs directly in the standard setting process, nor making clear how the P/ALDs connect to the cut scores
- Making sure the method fits the assessment design and intended summative uses (i.e, challenges with modifying or developing new standard setting methodologies)
- Training standard setting participants to think about assessment differently and train to different methodology

Deeper Dive: 6.4 - Reporting

The State reports its assessment results for all students assessed, and the reporting facilitates timely, appropriate, credible, and defensible interpretations and uses of those results by parents, educators, State officials, policymakers and other stakeholders, and the public.

1. **State-Level Reporting.** The state:
    a. Has clearly defined the intended purposes and uses of the state reports, and developed reports according to those purposes and uses
    b. Publically reports student achievement at each proficiency level and the percentage of students not tested, for all students as well as student group
    c. Provides guidance on the appropriate, and potentially inappropriate, uses of the reports
    d. Provides reports through a defined, timely distribution process

Deeper Dive: 6.4 - Reporting

The State reports its assessment results for all students assessed, and the reporting facilitates timely, appropriate, credible, and defensible interpretations and uses of those results by parents, educators, State officials, policymakers and other stakeholders, and the public.

2. **Individual- and Aggregate-Level Reporting**. The state provides interpretive, descriptive, and diagnostic individual-level student reports, and potentially aggregate reports, that:
   a. Are based on clearly defined the intended purposes and uses, and developed according to those purposes and uses
   b. Provide valid and reliable information about student achievement
   c. Report in terms of the grade-level academic standards, including in terms of P/ALDs
   d. Are useful in addressing academic needs
   e. Are accompanied by guidance on appropriate, and potentially inappropriate, uses
   f. Are available in alternate formats, if requested, or available in native languages, where practicable.
   g. Provided through a defined, timely distribution process

Deeper Dive: 6.4 - Reporting

Common Sources of Evidence:

1. **State Reporting**. Sample publicly accessible reports, interpretive guides, press releases, policies for distribution of individual reports, official communications to schools and districts.
2. **Individual and Aggregate Reporting**. Samples reports, interpretive guides, communications to districts, schools, and parents, instructions for retrieving reports/data files, report development processes, criteria, and stakeholder involvement, assistance provided for analyzing data files, documentation of locally developed tools and/or databases using test data for educational/instructional planning.

Common Pitfalls:

- Overly broad intended uses, beyond what the reports can support
- Reports that are not timely given the intended purposes and uses

Framing: Reporting

There is no requirement to report subscores.

There is a requirement to articulate the intended purposes and uses, and explain how the design of those reports and guidance materials supports those uses.

Framing: Reporting in terms of Use

Critical element 6.4 solicits evidence about the ways in which the program:

- provides information about "the specific academic needs of students" and
- "reports results for use in instruction".

The design of typical state assessment programs means that they are best suited **to *indirectly* influencing instruction**, e.g.,:

- Through the evaluation of curriculum
- Changes to instructional approaches in the upcoming year (e.g., corrections to what didn't work, tailoring to performance of incoming classes)

Programs should clarify the reasonable uses of the assessment results, and be **careful to avoid uses that are not supported.**

Unsupported uses can lead to problems.

Framing: Reporting in terms of Use

Critical element 6.4 solicits evidence about the ways in which the program:

- provides information about "the specific academic needs of students" and
- "reports results for use in instruction".

Some of the multiple approaches we examine are attempting to **increase direct instructional utility**.

Programs should clarify the reasonable uses of the assessment results, and be **careful to avoid uses that are not supported.**

Unsupported uses can lead to problems.

Framing: Reporting in terms of Use

- Critical element 6.4, both the element and the examples, solicit evidence about the ways in which reporting provides information about “the specific academic needs of students” and “reports results for use in instruction”.
- Typically state assessment indirectly supports instruction through actions that take place from year-to-year. Instead of claiming direct instructional utility instead be clear about what utility does exist, e.g.:
  - Through the evaluation of curriculum
  - Teacher reflection and corrections to their instructional approaches in the upcoming school year
  - Informing the upcoming year’s instruction by understanding the prior achievement of incoming students

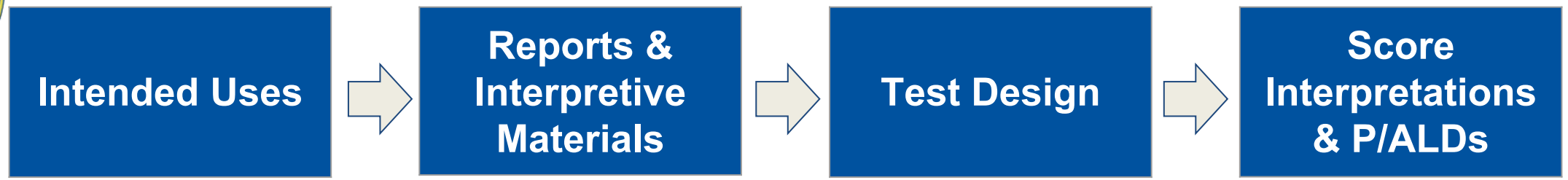
Framing: Backwards Design

Principled assessment design involves thinking starting with what we want the field to do with the results of assessment, **then designing backwards** from there.

Intended Uses	Reports & Interpretive Materials	Test Design	Score Interpretations & P/ALDs
What actions do we want specific stakeholders to take?	What reports and interpretive materials are need?	What kind of test should students take?	What do we want to say about what students know and can do?

Part of this logic is reflected in the subset of design and development activities shown above.

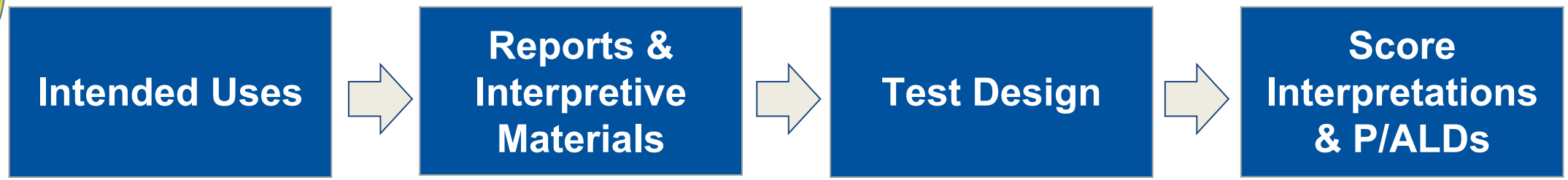
Framing: Backwards Design



- The above represents an ideal that isn't always implemented in practice, nor does doing so preclude a state from meeting peer review requirements. Often…
  - Development and implementation is messy and iterative
  - Parts of this backwards design process are skipped, assumed or otherwise not attended to
- Some design processes intentionally differ in terms of the order of some aspects of development (e.g., when P/ALDs are developed)

Peer review doesn't require a specific design approach. What is required is that intended use, and the supporting score interpretations, are well attended to!

Framing: Backwards Design



- For the multiple approaches we consider here, **the assessment program's purposes and uses are likely expanded beyond typical state summative assessment program**.
  - Clear articulation of the purpose and use is needed (drawing on evidence from CEs 2.1 and 3.1)
- The multiple approaches we explore here often involve changing:
  - What we want to say about what students know and can do, and thus changing the score interpretations and associated test design and P/ALDs
  - The uses assessments are put to, and thus the ways in which reports and guidance are developed and implemented

Limits of Peer Review

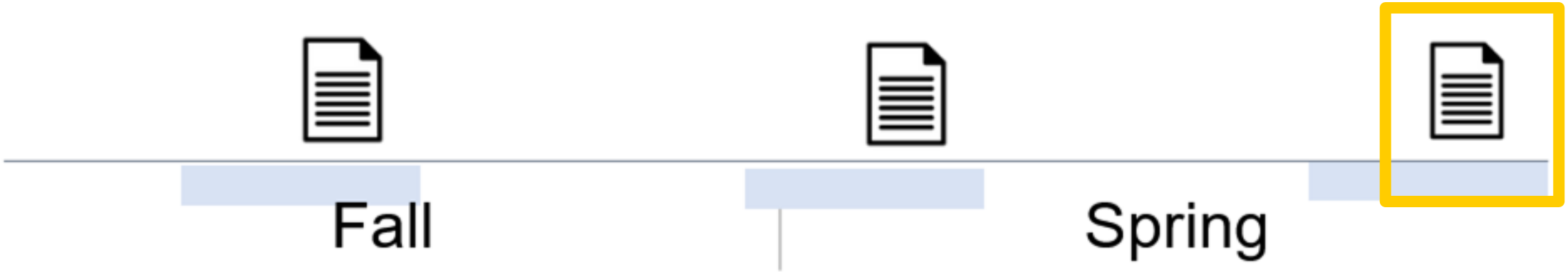
Peer review is concerned with the parts of the assessment program that are used **to produce annual determinations (i.e., scale scores and achievement level classifications)**:

> "… a State has the discretion to include in its assessment system components beyond the requirements of the ESEA, **which are not subject to assessment peer review**… A State also may include additional measures in its State assessment system, such as formative and interim assessments, which would not be subject to assessment peer review." (USDOE, 2018, p. 6, emphasis added)

Therefore, some parts of state assessment programs (e.g., some parts of a through-year assessment programs) fall **outside** of peer review.

Limits of Peer Review

- For example, if a program administers three assessments per year, but only the last is used to produce annual determinations, the preceding two assessments would **not** be submitted to peer review,
- Even if the first two were used to meet other purposes, e.g., so that "parents, teachers, principals, and administrators can interpret the results and address the specific academic needs of students" (USDOE, 2018, p. 71)



Limits of Peer Review

Grey areas include when results from assessments that **are not used to produce annual determiantions** are:

- Included on the individual or aggregate score reports for the assessment(s) produce annual determiantions**.**
- Used within the state's accountability system, e.g. using within-year growth

2. Issues and Options

2.A P/ALDs

Priorities and Choices in P/ALD Construction

- Defining assessable standards, and then translating those into achievement level descriptors involves **prioritizing what aspects of academic achievement** should be used to:
    1. Differentiate between levels of performance
    2. Communicate to the public
- This prioritization generally involves first defining policy P/ALDs, and then P/ALDs that are articulate the knowledge and skills from the learning standards, parsed across the different levels of performance (i.e., range P/ALDs). The development of these content referenced P/ALD can be done:
    1. Prior item development, or
    2. After item development

P/ALDs: Considering Multiple Approaches

What we want to say about what students know and can do?		
Through-Year	**Performance Assessment**	**Matrix Sampling**
Students have mastered a sufficient number of standards (e.g., DLM).	Students are able to apply their knowledge and skills to real world problems.	Across two years, students within a given school are able to demonstrate mastery on the depth and breadth of the content standards.
P/ALDs: designed around what a given number of standards master reflects in terms of content.	P/ALDs: designed around varying levels of application.	P/ALDs: designed around mastery, often in ways that match typical statewide assessment.

2.B Reporting

Expanded Reporting for Multiple Approaches

- Under these kinds of multiple approaches reporting will need to **expand.**
- This may include:
  - Reporting multiple times within the year
  - Including new metrics (e.g., comparisons across windows, including within-year growth)
  - New and increased interpretive documents, trainings and other supports to aid users
- These expansions include **individual student reports and similar reports**, but may also include expansions to the **state reporting system**.

Expanded Reporting due to Expanded Purposes and Uses

- **Expanded purposes and uses → expanded reporting**
  - Ideally, purposes and uses should be well described (e.g., in terms of a theory of action) so that reports can be connected to them
- **For any given purpose and use, there may need to be one or more tailored reports**, e.g,.
  - For administrators to use results to determine which teachers are most in need of intensive coaching, there may need to be multiple aggregated reports coupled with guidance materials

Expanded Reporting due to Expanded Purposes and Uses

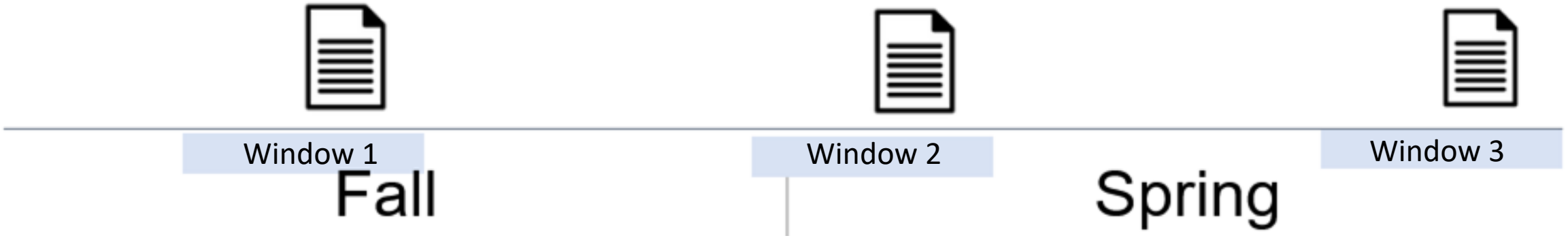
- Expanded purposes and uses → expanded reporting

Some of the implementations of these multiple approaches proposed are **expansive in their purposes and uses.**

This will need to be accompanied by **expansive reporting**.

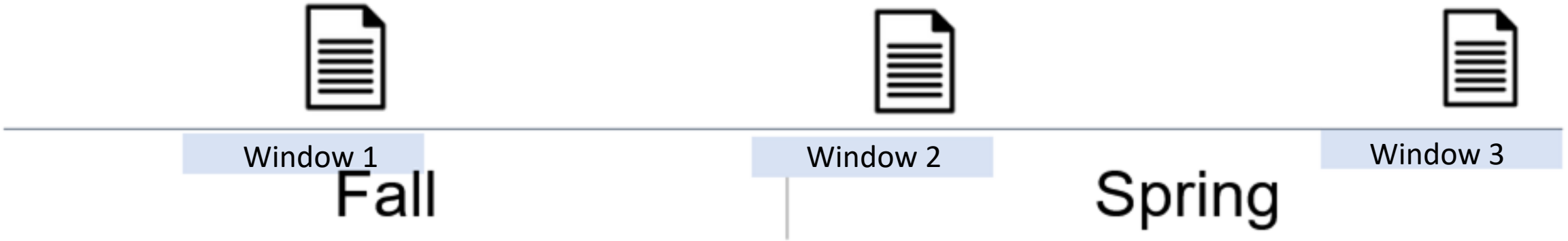
guidance materials

Considering Reporting for Within-Year Assessments



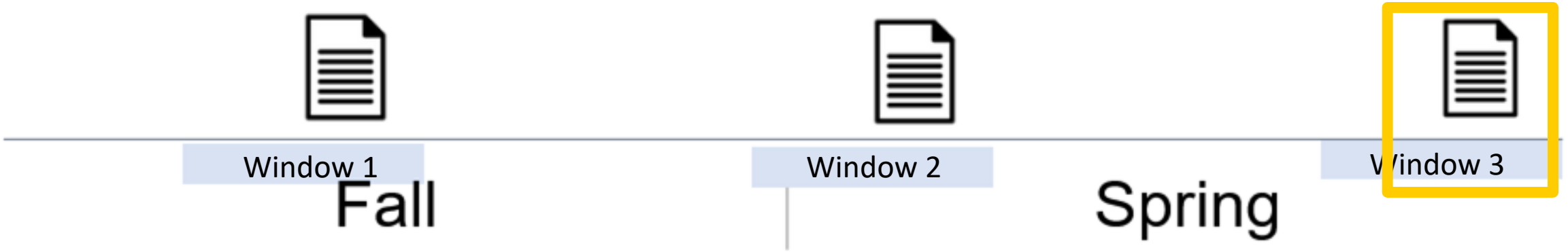
Under a design like the above, be it a through-year assessment or performance assessment, will involve:

- **Use.** Defining how the results from each administration are meant to be used (which may differ from window to window, as will the metrics. Designs may also differ from subject to subject)
- **Reports.** Developing and providing individual and aggregate reports for each window and supporting interpretive guidance
  - Considering coherence across the administrations (e.g., did well early on, but still not proficient)
- **Delivery.** Ensuring that the reports and guidance are timely and accessible



Additional challenges:

- **Instructional Utility**. Within-year reporting almost invariably is meant to support instruction directly, at least in part (e.g., teachers are meant to adjust their instruction in some way). **Defining and supporting these instructional actions continues to be extremely challenging.**
- **State Reporting**. A state may decide to incorporate results from the all of the windows within its state level reporting (e.g., dashboards, report cards), or just the last window.
- **Understanding**. More complex designs require more support to understand. Is sufficient support given to ensure understanding and avoid confusion?



- Again, peer review is limited to parts of the assessment program that are used to **produce annual determinations** (ie., scale scores and achievement levels) and the annual determinations themselves.
- Given this logic, if only the final assessment is used, only that assessment is the purview of peer review. However, this line may be blurred if prior assessments are used to provide information within the reporting on the final assessment (e.g., as subscores or across window growth)

Texas Through Year Assessment Pilot (TTAP)

- The Texas Educational Agency (TEA) has been engaged in extensive pilot development since 2019, including around reporting, with a to be determined implementation date.
  - This timeline has afford TEA with an opportunity to iterate with heavy user feedback loops
- Design is two administrations or “opportunities”, taking place in Fall and Winter), that are designed to be shortened forms of an also shortened end of year final opportunity
  - Each opportunity is meant to cover the “full scope” of the standards

Texas Through Year Assessment Pilot (TTAP)

- Current lessons learned and design decisions include:
  - A scoring model that doesn't penalize students for early low performance
  - Reports that contain both individual and group level predictions of later performance
  - Positioning our pilot such that it can meet peer review requirements and Texas statutes if/when we get there

Current areas of challenge include:

- Determining how to best display feedback in mainly static reports.
- How to show the instructional utility which is one of our pilot goals with the very short assessments and no subscore reporting.

NY's Performance-Based Components of Science Tests

- Two part test:
  - Multiple station-based, hands on science tasks
  - Written test
- Parts are combined to produce scale scores and performance levels, which are reported to parents; results (i.e., PLs) are aggregated for federal accountability.
- The performance tasks are scored against rubrics and available to teachers as individual task scores for instructional/evaluative uses.

NY's Performance-Based Learning and Assessment Networks (PLAN) Program

- Because performance tasks allow for more deliberate displays of knowledge and skill, the P/ALDs should be well connected to the scoring of the task(s) (e.g., rubric)

- More narrative reporting may be appropriate to describe students' level of achievement on these tasks (this also helps with required connection to instruction).

- Reported results are defensible in terms of purpose/use and the specific task(s) within the assessment.

Reporting: Matrix Sampling

- Matrix sampling content results in **reduced information at the student level**, and therefore may limit what can be reported at the student level
  - However, matrix sampling content can actually increase the amount of information reportable at aggregate levels, as the content standards could be covered in greater depth and breadth
- Individual level reporting will necessarily involve acknowledging that an individual student received a subset of the assessed content
  - Since not all students receive the same items, direct comparisons must be nuanced

3. Responding to Peer Review Requirements

General Guidance and Considerations

- The evidence needed for peer review may expand. Allow time for collecting and synthesizing the evidence.
- Clarity around intended uses will be key, as will be tying specific reports, interpretive materials and training to these uses.
  - Those submitting should sharpen and clarify both the current uses (e.g., to support district lead programmatic decision making as well as ESSA required school identification and support), as well as the new uses (e.g., supporting instruction).
  - Clearly delimint which purposes and uses fall under the purview of peer review, if applicable

P/ALDs

- Connect the design of the P/ALDs to the (1) the assessment program's intended purposes and uses and (2) assessment design (which should be articulated in 2.1 and 3.1)
- Clearly explain how the the P/ALDs were developed, including when, who was involved, etc.
  - Include all uses of the P/ALDs in item writing, standard setting, reporting, etc.
- If the P/ALDs were used to set cut scores, explain how they connect the learning standards with the resultant achievement standards.

Reporting

Clearly define the reports provided by the program and connect those to intended use. For example:

Report	Intended Audience	Intended Use
Individual Student Report	Parents	To inform parents of overall performance during the year, to encourage conversation with the student's teacher in the upcoming year about academic needs
Classroom Report	Teachers	To allow teachers to reflect on their instruction in the past year and adjust their instruction going into the current year
…		

Reporting

- Consider summarizing this work within a chapter on reporting within the technical report
  - Include blank templates
- In terms of development
  - Design backwards, if possible
  - Identify high-leverage, easy to access sources of feedback, if possible
  - Make the time and space for iterative reporting, if possible

4. 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