Session 1F: Achievement Level Descriptors (ALDs) and Score Reporting

Panelists: *Nathan Dadey, Meagan Karvonen, Zach Warner, Chris Rozunick*

[Session 1F Achievement Level Descriptors and Score Reporting slides 3–26]
Dr. Dadey provided an overview of the session, noting that ALDs and score reporting had been grouped together because they both affect the ways states communicate about assessment to the public and other audiences. He added that reporting often causes differences in opinion to surface too late in the process, so states should consider and consult on these issues during the design phase. The session covered CEs 6.2 (Achievement Level Standards Setting), 6.3 (Challenging and Aligned Academic Achievement Standards), and 6.4 (Reporting). These CEs focus on defining academic achievement of the standards and communicating that achievement to the field. They represent different components of the assessment design, implementation, and reporting process and have implications for almost all topics discussed in the conference.

After reviewing the relevant terminology, Dr. Dadey discussed the CEs in depth. For CE 6.2, the state must show that it used a technically sound method and process that involved panelists with appropriate experience and expertise for achievement standards setting. The proficiency level descriptors (PLDs) and ALDs and achievement standards (i.e., cut scores) must be reasonable and defensible. The cut scores must be reliable and distinguish between the levels described by the PLDs and ALDs. Evidence can include the PLD or ALD development process/script, the standards setting process/script, relevant training materials and reports, TAC minutes, and participant rating summaries. States might also describe the panelist characteristics and selection criteria. Common pitfalls include having panelists who are not representative of educators who understand all students (e.g., having no special education teachers or teachers who work with English learners), not providing enough detail or justification for the standards setting process, and not clearly explaining the process during training.

For CE 6.3, the state must show that its academic achievement standards are challenging and aligned with its academic content standards and the knowledge and skills necessary for success in college and the workforce. Evidence may include a summary or report of the ALD development process or an expert opinion on AA-AAAS. States may also provide documentation of the vertical articulation processes for PLDs, ALDs, and standards setting; summaries of students in each performance level; item mapping studies; comparisons with external benchmarks; TAC minutes; a crosswalk between ALDs and content standards; and alignment analyses. Common pitfalls when submitting documentation to peer review include challenges deriving achievement level labels that are meaningful, support intended interpretations, and are agreed upon by all stakeholders.

For CE 6.4, the state must report its assessment results for all students assessed. The state must also demonstrate that 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. For state-level reporting, it must demonstrate clear intended purposes and uses of these reports and develop reports accordingly. States should publicly

report student achievement at each proficiency level (and the percentage of students not tested) for all students and student groups. The state should also provide guidance on the appropriate (and potentially inappropriate) uses of the reports and provide them through a defined, timely distribution process. For individual- and aggregate-level reporting, the state must provide interpretive, descriptive, and diagnostic individual-level student reports. It may provide aggregate reports that, among other things, provide valid and reliable information about student achievement, are useful in addressing academic needs, and report the grade-level academic standards (e.g., PLDs and ALDs). As evidence of meeting this CE, the state may provide sample publicly accessible reports, interpretive guides, official communications to districts and individual schools, and documentation of locally developed tools using test data for instructional planning. Common pitfalls for reporting include overly broad intended uses and reports that are not timely, given the intended purposes and uses.

Turing to how states can frame reporting, Dr. Dadey emphasized that there is no requirement to report sub-scores. He acknowledged that the field may prefer that states report sub-scores. States are required to articulate the intended purposes and uses and to explain how the report design and guidance materials support those uses. To frame reporting in terms of use, states may meet CE 6.4 by submitting evidence that the assessment and reporting program provides information about "the specific academic needs of students" and reports results for use in instruction. Typically, state assessment programs are best suited to indirectly influencing instruction. However, some of the multiple approaches discussed at the conference aim to increase direct instructional utility. It is important for states to clarify the reasonable uses of the assessment results and to avoid uses that are not supported.

Dr. Dadey emphasized that ideally, principled assessment design involves considering what the field will do with the results of assessment and then designing backward—from intended uses to reports and interpretive materials through test design to score interpretations and PLDs and ALDs. States are likely to fall short of this ideal but can still meet peer review requirements, as long as they attend to intended use and supporting score interpretations. Multiple approaches will most likely expand the assessment program's purposes and uses beyond those of the typical summative assessment. States using these approaches must clearly articulate the purpose and use, drawing on evidence from CEs 2.1 and 3.1. They may need to change the uses of assessments, score interpretations, test design, and PLDs and ALDs, as well as the ways in which reports and guidance are developed and implemented.

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). Therefore, some parts of state assessment programs fall outside 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. Gray areas include when results from assessments that are not used to produce annual determinations are included on the individual or aggregate score reports for the assessments and used within the state's accountability system.

[Session 1F Achievement Level Descriptors and Score Reporting slides 27–34]
Dr. Warner focused on the choices that states can make regarding PLD and ALD construction, which depend on the levels of academic achievement (e.g., standards and proficiency) that are valued. Defining assessable standards and then translating those into ALDs involves prioritizing the aspects of academic achievement that should be used to differentiate between levels of performance and communicate to the public. This prioritization generally involves first defining policy PLDs and ALDs and then ones that articulate the knowledge and skills from the learning standards, parsed across the different levels of performance (i.e., the range of PLDs and ALDs). He noted that ESSA covers three levels of performance, but some states target more. Each level is defined and linked to policy. It is best to go through the validity argument, design decisions, and reporting considerations prior to item development.

In TYA, students show that they have mastered a sufficient number of standards. PLDs and ALDS are designed around what mastery of a given number of standards reflects in terms of content. In performance assessment, students apply their knowledge and demonstrate observable skills to real-world problems. PLDs and ALDs are designed around varying levels of application. In matrix sampling, students in a school have the opportunity to demonstrate mastery on the depth and breadth of the content standards across two years. PLDs and ALDs are designed around mastery, often in ways that match typical statewide assessment. These multiple approaches will expand reporting because of complex or more frequent assessment. The state's intended purposes and uses may also expand. Therefore, the state may need to report multiple times within the year, include new metrics, and provide new and increased interpretive documents, trainings, and other supports to assist users. Expansions include individual student reports but may also include changes to the state reporting system.

[Session 1F Achievement Level Descriptors and Score Reporting slides 35–39]
Ms. Rozunick discussed reporting for within-year assessments (e.g., a state that offers three windows for testing and reporting for each). The state will need to define the intended use for each administration and metrics (which may differ from window to window). Designs may differ among subject areas. It will develop and provide individual and aggregate reports for each window, along with interpretive guidance. Important considerations for within-year assessments also include the coherence across the administrations and the provision of timely and accessible reports and guidance. Challenges to states offering within-year assessments may include defining and supporting instructional actions based on assessments. Incorporating the results from all the windows within state-level reporting may be difficult. Finally, stakeholders will need sufficient support to understand more complex assessment designs. If only the final assessment is used to produce annual determinations, then peer review only covers that assessment. However, this line may be blurred if prior assessments are used to provide information within the reporting on the final assessment (e.g., as sub-scores or across window growth).

TTAP provides an example, as it offers two test administrations (in the fall and winter) that are designed to be shortened forms of an abbreviated end-of-year assessment. Texas decided on a design that includes a scoring model that does not penalize students for early low performance

and reports that contain both individual- and group-level predictions of later performance. The state is positioning this pilot program so that it can meet peer review requirements and Texas statutes in the future. TTAP has presented some challenges, such as determining how to best display feedback in mainly static reports and how to show the instructional utility with short assessments and no sub-score reporting.

[Session 1F Achievement Level Descriptors and Score Reporting slides 40–45]
Dr. Warner explained that in New York's science assessment, the performance-based component and written test are combined to produce scale scores and performance levels, which are reported to parents. The results are aggregated for federal accountability. The performance tasks are scored against rubrics and available to teachers as individual task scores for instructional uses. New York was awarded funds from the CGSA program to expand performance-based learning and assessment. Because performance tasks allow for more deliberate displays of knowledge and skill, the PLDs and ALDs should be well connected to the scoring of the tasks (e.g., the rubric). More narrative reporting may be appropriate to describe students' level of achievement on these tasks, which also helps with the required connection to instruction. The report results are defensible in terms of purpose and use and the specific task(s) included within the assessment. PLDs can have district- and school-level applications.

Content derived from matrix sampling results in reduced information at the student level and therefore may limit what can be reported. However, matrix sampling content can actually increase the amount of information reportable at aggregate levels, as the content standards can 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. Because not all students receive the same items, direct comparisons must be nuanced.

When preparing submissions for peer review, states using performance-based assessments may need to provide more evidence than usual and need to allow time for collecting and synthesizing it. They should be clear about the intended uses of assessment and link specific reports, interpretive materials, and training to these uses. Dr. Warner stressed the importance of creating a cohesive and coherent but easy-to-follow argument about the assessment design with evidence to peer reviewers. States should connect the design of the PLDs and ALDs to the assessment program's intended purposes and uses and overall design (which should be articulated in CEs 2.1 and 3.1). They should explain in detail how the PLDs and ALDs were developed. If the PLDs and ALDs were used to set cut scores, then the state needs to explain how learning standards are connected with the resultant achievement standards.

[Session 1F Achievement Level Descriptors and Score Reporting slides 46–47]
Dr. Dadey reviewed some final considerations for reporting. States should clearly define the reports provided by the assessment program and connect them to the intended use. For example, individual student reports are received by parents so they can be informed of their child's overall performance during the year and encourage conversation with the student's teacher(s) about academic needs in the upcoming year. States should consider summarizing their work (e.g., in blank templates) within a chapter on reporting within the technical report.

To develop reports, use backward design and identify high-leverage, easy-to-access sources of feedback.

Questions and Comments

Does New York investigate the instructional utility of PLDs and ALDs as part of its argument for performance-based assessment?

Dr. Warner responded that New York has made the case that the interpretation of PLDs has utility for instruction. The state consults teachers to determine how the skills associated with learning standards might be assessed when developing PLDs.

In the past, my state has not submitted evidence regarding efforts related to data literacy for peer review. Should we submit this information?

Dr. Peasley remarked that the state probably could support some CEs in particular contexts with evidence of data literacy efforts but that it would be best to have a conversation with ED.

Does content-based matrix sampling conflict with educators' desire for actionable student-level data? If so, how can we resolve this issue?

Dr. Warner commented that matrix sampling can conflict with the desire for actionable student-level data, but states need to discuss this trade-off in the context of assessment design. With less testing, the state obtains less information. States that want a great deal of individual student information might want to select a different type of assessment system.