

AUGUST
2017

CONSIDERATIONS FOR ADOPTING AND IMPLEMENTING INNOVATIVE ASSESSMENT SYSTEMS

Submitted to: The Colorado Department of Education

By: Benjamin R. Shear, Elena Diaz-Bilello and Rajendra Chattergoon
Center for Assessment, Design, Research and Evaluation (CADRE)
University of Colorado Boulder





ACKNOWLEDGEMENTS

We thank Alyssa Pearson, Joyce Zurkowski and Marie Huchton at the Colorado Department of Education (CDE) for their valuable input and careful review of this report. We are also grateful to Derek Briggs at the Center for Assessment, Design, Research and Evaluation, University of Colorado Boulder for his helpful comments and feedback on an earlier draft of this report.

ABOUT CADRE

The Center for Assessment, Design, Research and Evaluation (**CADRE**) is housed in the [School of Education at the University of Colorado Boulder](#). The mission of CADRE is to produce generalizable knowledge that improves the ability to assess student learning and to evaluate programs and methods that may have an effect on this learning. Projects undertaken by CADRE staff represent a collaboration with the ongoing activities in the School of Education, the University, and the broader national and international community of scholars and stakeholders involved in educational assessment and evaluation.

PLEASE DIRECT ANY QUESTIONS ABOUT THIS REPORT TO:
Benjamin.Shear@Colorado.Edu

EXECUTIVE SUMMARY

The federal Every Student Succeeds Act (ESSA) will allow up to seven states the flexibility to pilot “innovative assessment systems” to fulfill accountability testing requirements under the Assessment and Accountability Demonstration Authority (the “demonstration pilot”). Although the demonstration pilot allows greater flexibility by not requiring use of a single, year-end statewide test for all students, any system submitted under the pilot must still satisfy federal and state reporting requirements for the same grades and subjects that current legislation requires. This report provides background information to assist the Colorado Department of Education (CDE) in fulfilling the requirement of House Bill 16-1234 (HB 16-1234) to consider assessment system designs that use district-selected or district-created assessments for accountability reporting and that can also be submitted for the demonstration pilot.

The report first describes five assessment systems being developed in other states and in Colorado. The two systems being developed in other states (New Hampshire and Arizona) are anticipated to be used for demonstration pilot applications in those states. The three systems from Colorado provide an indication of the types of assessment systems valued by districts, but are not necessarily designed to fulfill the demonstration pilot requirements. The report then explains the concepts of validity, reliability, comparability, alignment, and fairness, which are identified in the demonstration pilot and HB 16-1234 legislation as key assessment quality criteria to consider. The explanations provided in the report draw on published professional standards in educational measurement. The report also identifies two nationally recognized sets of assessment criteria that could inform the CDE’s work: one developed by the Chief Council of State School Officers (CCSSO) and the other by the US Department of Education. The report highlights potential strengths and weaknesses (including approximate costs) of the five example assessment systems relative to the various criteria. Although none of the assessment system examples considered in the report appear to fully meet the requirements of both the demonstration pilot and Colorado statutes yet, they are used to illustrate how the criteria can be used to evaluate potential assessment system designs. They are also used to highlight the extensive work and resources that will be required to design, implement, and evaluate assessment systems (e.g., the performance based assessment system used in New Hampshire).

While the demonstration pilot legislation allows states greater flexibility in the design of accountability testing systems, state and federal reporting requirements severely restrict the amount of flexibility states will have in designing these “innovative assessment systems.” The three primary constraints identified in this report are:

1. To produce an annual, summative achievement determination for every student in federally required grades and subjects that is comparable statewide.



2. That all students, including students with disabilities and English language learners, be assessed in appropriate and fair ways.
3. That assessment results can be used to produce student growth percentiles for every student in required grades and subjects.

Requirement (3) is based on Colorado's Senate Bill 191 (SB-191), requiring an assessment system that can produce scores appropriate for use with the Colorado Growth Model. Designing a system that meets these requirements and accomplishes other aims of the state, for example to reduce overall state-mandated testing or to improve the instructional relevance of accountability testing, will entail making some difficult tradeoffs. One apparent way to serve both aims is by designing a system that uses the same assessments for multiple purposes. Both New Hampshire and Arizona plan to use district-created or district-selected assessments administered throughout the school-year in place of a year-end statewide test. The goal is for these assessments to fulfill accountability reporting requirements and to provide information that is useful locally within districts, but these dual purposes raise challenges. It remains unclear whether these systems will be able to adequately satisfy the demonstration pilot requirements that all students are assessed in a manner that is fair and comparable across the state, or whether they could be used to satisfy the Colorado Growth Model requirements. It is also unclear whether using these assessments for high-stakes accountability purposes will have adverse consequences on their utility for other purposes, such as providing instructionally relevant information to teachers.

Designing an assessment system must begin by clearly identifying the aims the system will accomplish. Ideally this would drive the design of a balanced assessment system, in which different forms of assessment most appropriate to each intended use are coordinated across the system. Implementing such a system is challenging, given the autonomy of districts to design their own curricula, the desire to reduce the overall amount of state-mandated testing, and the existing federal and state legislative requirements. The fundamental challenge in designing an assessment system under the demonstration pilot seems to be this: the flexibility built into the demonstration pilot – that the same test need not be administered to all students to make annual achievement determinations – is undermined by the requirement for such determinations to be comparable for all students (including those taking the current statewide test). Many innovative assessment system designs that could improve the authenticity of assessment tasks or reduce testing time, such as greater use of performance tasks or matrix sampling, are unlikely to satisfy the constraints placed by existing legislation.



TABLE OF CONTENTS

Introduction	6
Examples of Assessment Systems	8
State Examples: A Performance Based Assessment System and an Interim Assessment System	10
• <i>Locally Developed Performance Based Assessment System Example: New Hampshire</i>	10
• <i>Locally Selected Interim Assessment System Example: Arizona</i>	11
Local Assessment System Examples	12
• <i>Common, Integrated Interim and Summative Assessment Example: Cherry Creek</i>	13
• <i>Locally Selected Assessment System Example: S-CAP Districts</i>	14
• <i>Combination of Common Interim and Locally Selected Assessments Example: Westminster Public Schools</i>	15
Assessment Quality Criteria	16
Assessment Systems	17
Definitions of Quality Criteria	18
• <i>Reliability</i>	20
• <i>Comparability</i>	21
• <i>Alignment</i>	22
• <i>Fairness</i>	23
• <i>Additional Criteria</i>	24
Challenges for Meeting Quality Criteria in Innovative Assessment System Examples	24
Existing Assessment Quality Frameworks	27
• <i>CCSSO Criteria</i>	27
• <i>Federal Peer Review Critical Elements</i>	27
Conclusion	29
References	32
Appendix A	35
Appendix B	38
Appendix C	46
Appendix D	48
Appendix E	49
Appendix F	83



INTRODUCTION

The Colorado Department of Education (CDE) commissioned the Center for Assessment, Design, Research and Evaluation (CADRE) at the University of Colorado Boulder to compile a report describing assessment systems and assessment quality criteria to inform possible changes to state accountability testing. House Bill 16-1234 (HB 16-1234) specifically calls for an investigation into “alternative summative assessment models” that would provide “valid, reliable and comparable” data, with consideration for determining whether “the assessments are suitable for the state accountability system” (pg. 3). HB 16-1234 also requires CDE to apply for the United States Department of Education (USED) Innovative Assessment and Accountability Demonstration Authority pilot (hereafter referred to as, “demonstration pilot”) under the Every Student Succeeds Act (ESSA), as practicable. The demonstration pilot will grant up to seven states the flexibility to pilot accountability assessment systems that do not necessarily include administering a single, statewide year-end summative assessment in grades 3-8 and once in high school to inform school accountability decisions, as is currently required. Under the demonstration pilot, states may begin by piloting the system in a small number of districts, replacing the regular schedule and expectations for administering the summative state-wide assessment only in pilot districts. Additionally, there must also be a plan for scaling the system up to the statewide level over time.

The purpose of this report is to: describe examples of assessment systems currently being developed or implemented that could serve as a basis for a demonstration pilot application, highlight challenges these systems are likely to face, and describe relevant issues for the state to consider when submitting a proposal for the demonstration pilot

Neither HB 16-1234 nor the demonstration authority legislation explicitly define what constitutes an “alternative assessment model” or “innovative assessment system.” The federal legislation does, however, state that innovative assessment systems eligible for consideration are assessment systems that may include either:

§1204(a)(1): “competency based assessments, instructionally embedded assessments, interim assessments, cumulative year-end or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments...”

or:

§1204(a)(2): “assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs.”



Although the term “assessment system” is not explicitly defined, the implied usage (and the usage we adopt here) is that an “assessment system” refers to the overall collection of assessments that is used to make annual summative achievement determinations (i.e., to make judgments about proficiency) for each student in the required grades and subjects. Although the demonstration authority legislation allows states greater flexibility in the design and use of assessments in their accountability systems, the legislation still makes two critical requirements that we will discuss at length in this report. First, states must continue to report annual achievement determinations for all students in mathematics and English Language Arts in grades 3-8 and in one level in high school; annual achievement determinations are also required for one grade in each level (elementary, middle and high school) for science. Second, although these determinations need not be based on a single, statewide summative test,¹ the determinations must be deemed “comparable” for all students at a statewide level. Briefly, this means that even if districts administer different tests as part of the demonstration pilot, a state must provide evidence that the annual achievement determinations across districts still have the same meaning across districts. We address this comparability requirement later in this report, as it represents a critical piece to consider when designing an assessment system for the demonstration pilot.

This report is organized as follows. First we provide a brief overview of two different types of “innovative assessment systems” that are expected to be used for demonstration pilot applications in New Hampshire and Arizona. We then describe examples of local assessment systems currently being developed in three Colorado sites. For each site we describe the goals of the assessment system and the specific assessments the sites are either already using or plan to use. We then review key assessment quality criteria that should be considered when evaluating assessments used in the demonstration pilot, and discuss how these criteria relate to the assessment system examples. Finally, the report concludes with a summary of our findings and a discussion of additional considerations relevant to designing an assessment system for the demonstration pilot.



¹ Consistent with the literature on testing and assessment, we use the terms “test” and “assessment” interchangeably throughout this report, unless specified otherwise.

EXAMPLES OF ASSESSMENT SYSTEMS

In this section we describe five assessment systems that use different combinations of assessments to make judgments about student proficiency. The two state examples (New Hampshire and Arizona) were selected because these states have declared their intent to apply for the demonstration pilot and are instituting systems that represent two very different approaches: a system based on locally developed performance tasks (New Hampshire) and a system based on pre-existing interim assessments selected independently by districts in Arizona. These two sites were also selected because the design of both systems was informed by a desire to promote local control at the district level.

We then turn to describing examples of assessment systems being developed in three Colorado sites: Cherry Creek School District, the Student-Centered Accountability Project (or “S-CAP”) school districts and Westminster Public Schools. The three Colorado sites were selected for this report because CDE staff identified them as sites that were exploring alternative assessment systems. Cherry Creek uses an integrated system of interim and summative assessments that are common across all schools. The S-CAP allows each member district to independently develop or adopt assessments to evaluate student learning. Westminster Public Schools requires all schools to administer a common interim assessment, but allows schools to independently develop or adopt assessments used to make end-of-year competency determinations. At this time, all of these sites still administer the statewide Colorado Measures of Academic Success (CMAS) assessments in the usual grades and subjects, as is required by current legislation.

Appendix A provides approximate per pupil cost estimates associated with different aspects of implementing and/or developing the types of assessments discussed in the examples. It is critical to note that the per pupil cost estimates provided in Appendix A cannot be directly compared because the estimates are based on different assumptions and are also based on data for the year that a given site was willing to provide or share with us. The cost estimates are provided to illustrate the type of expenses that could be incurred for different systems, but likely underestimate the total cost due to additional indirect expenses that are difficult to quantify precisely. For example, these costs do not take into consideration the data systems that would need to be put into place at the state level should any of these approaches be used for state-wide accountability.

Table 1 presents a brief description of the different types of assessment systems and types of assessments used in each example described in this report.

Table 1. Types of assessments used in each site reviewed.

Assessment System Description	Example Site	Cumulative year-end assessments (for all grades)	Instructionally-embedded assessments	Interim assessments	Performance-based assessments
Locally developed performance tasks	New Hampshire*		✓		✓
Locally selected interim assessments	Arizona†			✓	
Common, integrated interim and summative assessments	Cherry Creek‡	✓		✓	
Locally selected assessments (any form)	S-CAP Districts‡	✓	✓	✓	✓
Combination of common interim and locally selected assessments	Westminster‡	✓	✓	✓	

* Restricted to districts participating in the Performance Assessment of Competency Education (PACE) pilot in New Hampshire

† We focus only on the grades 3-8 system proposed for Arizona.

‡ In Westminster, Cherry Creek and the S-CAP districts, the statewide year-end assessments are still included in the table because the state assessments will continue to be administered and used by those sites in all federally-required tested grades for the foreseeable future.

Four of the five examples are considering the possibility of breaking from the current conventional role of using cumulative statewide year-end assessments in grades 3-8 and in one grade in high school to inform annual achievement determinations for school accountability. New Hampshire uses locally developed performance tasks as the primary component in their assessment system, and administers the cumulative state year-end assessments in only one grade for each level (i.e., elementary, middle and high school) in participating districts. Over time, New Hampshire intends to scale up this approach and have all districts use the performance assessments and state year-end assessments in the same manner. The proposed assessment system in Arizona for grades 3-8 would end the use of any statewide year-end assessments. The Cherry Creek School District and the districts participating in the S-CAP are exploring the possibility of phasing out use of the statewide year-end assessment. At present, these districts will continue to use the statewide assessment as required by law, and a concrete plan or timeline for the phase-out of the statewide assessment remains uncertain. Westminster Public Schools has not determined what role the statewide year-end assessment would ideally play in its system.



STATE EXAMPLES: A PERFORMANCE BASED ASSESSMENT SYSTEM AND AN INTERIM ASSESSMENT SYSTEM

After two years of development and preparation, New Hampshire's Performance Assessment of Competency Education (PACE) model is currently being piloted in eight of the state's 166 school districts and in one charter school. Arizona's proposed model for accountability testing has not yet been implemented, but we describe their proposed plan for grades 3-8. We do not describe Arizona's high school model, which will employ nationally recognized assessments to make annual determinations in one or more grade levels, and will likely not require an application for the demonstration pilot.

Locally Developed Performance Based Assessment System Example: New Hampshire

The PACE model used in New Hampshire was developed around a theory of action based on fostering deeper learning in students and improving instructional practices. The state selected performance assessments to serve as the centerpiece of the model, with the belief that these types of assessments can better elicit information about student reasoning and understanding (NHDOE, 2015). At the end of the 2016-17 school year, the state will conclude a three-year pilot approved by USED separately from the demonstration pilot. The goal of the initial pilot was to build capacity in participating districts to implement the PACE system. For districts participating in the PACE project, the Smarter Balanced assessment (New Hampshire's current statewide year-end summative test) is administered each year to all students in only one grade for each level (elementary, middle and high school).² Even for grades in which students take the Smarter Balanced assessment, these results are combined with other assessments completed throughout the school year to make annual competency determinations for each student.

Teachers play a key role in the PACE project by developing, administering, and scoring the performance tasks used to inform competency judgments for each content area and grade level. Currently, all PACE districts must use one common performance task in each grade and content area that does not use the statewide assessment. These common tasks are used as a score calibration tool to ensure that teachers across districts use the same expectations for scoring and rating the performance of their students, including on other locally developed tasks. The belief is that over time, meetings structured around common performance tasks will ensure that proficiency expectations for students across different school districts will become consistent, despite differences in the locally developed assessments used in those districts. Teachers use a quality assessment review tool developed by external consultants to determine whether the common tasks and



² Per the New Hampshire Department of Education, "Smarter Balanced is administered in grade 3 (English language arts), 4 (math), and grade 8 for both ELA and math. The SAT is administered to all grade 11 students. In other words, "statewide" assessments are administered in only 6 grades/subjects and local assessments in 17" (pg. 4).

other local assessments used in each district meet high quality expectations. A copy of this assessment review tool is included in Appendix B. The local assessments are also reviewed by external consultants who provide feedback used to improve the quality of the assessments. This process highlights the essential role that the state department of education and external consultants must play in facilitating the collaborative work of teachers in the PACE project.

To make year-end performance level achievement determinations for each student, teachers in the participating districts created achievement level descriptors for each content area and grade level. These descriptions were based on the Smarter Balanced achievement level descriptions. Teachers then evaluate a cumulative body of each student's work from the school year (including the locally developed assessments) to determine each student's achievement level. While the locally developed performance assessments comprise the majority of this body of evidence, the body of evidence also includes either the single common performance task or the statewide assessment, depending upon the grade and subject area.

When USED granted permission to New Hampshire to pilot this competency-based assessment system starting in the 2014-15 school year, the New Hampshire Department of Education indicated that the annual proficiency determinations from the PACE assessments would be used for school accountability after the final year of the three-year pilot. As a result, PACE districts and schools plan to use the student proficiency results based on the PACE system for school accountability purposes beginning with the 2017-18 school year.

Figure 1 in Appendix A presents the estimated costs associated with implementing the PACE system in New Hampshire in year one of the three-year pilot granted by the USED. These estimates only reflect costs documented in the 2014-2015 school year for the original pilot districts. These estimates do not include logistical costs associated with implementing the pilot, such as compiling and reporting scores in each district.

Locally Selected Interim Assessment System Example: Arizona

House Bill 2016-2544 in Arizona calls for the state board of education to adopt a "menu of locally procured achievement assessments to measure pupil achievement of the state academic standards." This menu approach would be applied to grades 9-12 during the 2017-18 school year and would be extended to grades 3-8 in the 2018-19 school year. Unlike New Hampshire, no claim is made about the desire for this model to foster deeper learning in students, although the assessment system proposed for grades 3-8 in Arizona is intended to aid instruction. In brief, the proposed system would replace use of the statewide year-end assessment to make annual achievement determinations for each student with scores from interim assessments selected by each district

independently. A key claim of this menu approach is that overall testing time would be reduced because interim assessments already being used by districts would now also be used for accountability purposes.

According to Irene Hunting (former Deputy Associate Superintendent for Assessment at Arizona's Department of Education) this menu approach is intended to provide flexibility for participating districts to determine which assessments (i.e., external, vendor-developed assessments) can best evaluate student learning (personal communication, September 12, 2016). Under Arizona's proposed system, the onus of ensuring that any interim assessment used for the pilot meets federal peer review standards would fall on the district or the test vendor. The state would primarily serve in a support role for participating districts. The state's role would include ensuring that assessment data are collected in a timely fashion for accountability purposes and submitting evidence on behalf of vendors to demonstrate that a given assessment provides high-quality, comparable data that can be used to inform school accountability ratings. At present, there is no clear plan in place regarding how the state will ensure that results from different assessments used in different districts are comparable.

Participating districts, not the state, would bear all costs associated with administering, scoring and reporting the test results. Districts would also bear the cost of submitting their selected assessment for peer review, although these costs may potentially be shared by other districts using the same assessment or by the test vendor. Although no cost estimates can be provided by the state for supporting this district-based approach, we provide annual per-pupil cost estimates for purchasing the Measures of Academic Progress (MAP) interim assessments from the Northwest Evaluation Association (NWEA), which are widely used in Arizona. The estimates were shared with us by the Cherry Creek School District in Colorado. Estimated costs are located in Figure 2 of Appendix A. These cost estimates do not include the costs associated with determining inter-district comparability, going through federal peer review, training teachers to make use of the assessments, or conducting analyses and generating reports for statewide accountability determinations.

LOCAL ASSESSMENT SYSTEM EXAMPLES

This section presents a brief overview of three assessment systems being used to make proficiency or competency determinations in Colorado school districts. District personnel at the three sites deem these systems to be at the early stages of development and would caution against stating that the information presented below reflects final decisions for either the assessment or accountability designs. Cherry Creek School District and the S-CAP group are focused on building local accountability systems that could eventually replace the state's school performance framework (SPF). The third site, Westminster Public Schools, does not intend to



create a local accountability system separate from the state's SPF. It is important to note here that none of these sites have conducted an alignment study to determine the extent to which their selected assessments are aligned with state standards as stipulated by the demonstration pilot and by HB 16-1234. This is partly because all sites intend to continue using the annual state-wide summative assessments for the foreseeable future and therefore would not need to conduct an alignment study at this time. However, if one or more of these sites decides to phase out the annual state-wide summative assessment, then an alignment study would be required. We discuss alignment in more detail in a later section of this report.

Common, Integrated Interim and Summative Assessment Example: Cherry Creek

The Cherry Creek School District (CCSD) has used a suite of integrated interim and summative assessments (the ACT Aspire system) for several years and is currently planning to use these assessments as the primary component in their own school accountability system in grades 4-9. The pre-ACT assessment is administered to grade 10 students and all grade 11 students take the ACT. According to Dr. Judy Skupa, Assistant Superintendent for Performance Improvement in CCSD, the rationale for prioritizing use of ACT assessments over the statewide assessments stems from three concerns: 1) the amount of time required for students to take the state assessments, 2) the length of time it takes for results from the state assessments to return to districts, and 3) the continuously changing state assessment landscape, which appears to be influenced by political motivations (personal communication, April 21, 2016). The ACT Aspire system offers both interim and summative assessments in English, Mathematics, Reading, Science and Writing, in grades 4-11. According to Dr. Skupa, CCSD believes these assessments would be deemed as more "valid" than the year-end state assessments in the eyes of district-based stakeholders, parents, and the larger community because these assessments form an "aligned" system to evaluate college readiness goals and provide a stable and consistent source of data for all students in federally required tested grades.

Presently, CCSD staff is working with school and community-based stakeholders to design the local accountability system. Although the ACT Aspire assessments will likely play a prominent role in the accountability system, CCSD may also consider incorporating additional assessments valued by school staff to evaluate student performance.

The per pupil cost estimates for the ACT Aspire assessments are located in Figure 3 of Appendix A. Again, these estimates only represent the direct, per-pupil costs of subscribing to the ACT Aspire system and do not include additional costs associated with the state data and accountability systems.



Locally Selected Assessment System Example: S-CAP Districts

The Student-Centered Accountability Project (S-CAP) has a membership of five rural school districts.³ According to the S-CAP website, this project has the purpose of “whole child” accountability through continuous improvement of the educational system. To accomplish this, the S-CAP plans to use multiple measures to evaluate the success of students and to evaluate the capacity of the system. The following four components comprise the requirements for S-CAP members to include in the accountability system:

1. Student Achievement: each district must pick two measures of student achievement to evaluate student performance or “academics” across schools.
2. Learning Dispositions: all districts will use the same tool to evaluate student learning dispositions (i.e., engagement and mindsets).
3. Professional Culture: all districts will use the same school quality review tool to evaluate the professional culture across schools.
4. Resource Allocation: all districts must also evaluate finances, infrastructure and facilities/safety, and family and community.

Under the S-CAP local accountability system, items (3) and (4) represent the “inputs” that influence the results or “outputs” captured by items (1) and (2). The S-CAP allows participating districts to select their own measures of student learning for item (1), and these can differ across districts. Buena Vista, for example, uses their own teacher-developed assessments to measure student learning, whereas Merino Buffalo uses the NWEA MAP assessments to measure student learning.

Aside from the student achievement measures, which are already in use by all S-CAP districts, the other components of the S-CAP system are still being developed or refined. Since the S-CAP members are currently in the process of selecting or developing a common instrument to evaluate learning dispositions, this area will not be evaluated until districts have the opportunity to pilot and learn from those results. The S-CAP districts are also continuing to refine both the instrument and the review process used to assess professional culture and resource allocation.

Figure 4 in Appendix A presents information on costs associated with developing the S-CAP accountability system. The estimates provided in Figure 2 for Arizona districts using MAP would apply to S-CAP districts using the MAP or similar interim assessments, whereas estimated costs associated with teacher-developed assessments are more difficult to determine.



³ S-CAP districts are: Buena Vista, Buffalo Merino, Kit Carson, La Veta, and Monte Vista.

Combination of Common Interim and Locally Selected Assessments Example: Westminster Public Schools

The competency-based education model used in Westminster Public Schools (WPS) defines the curricular, instructional, and assessment opportunities offered to students in the district. According to Dr. Oliver Grenham, Chief Education Officer of WPS, the assessments used in the district support curricular and instructional goals linked to competencies developed for each content area and grade (personal communication, November 1, 2016). Although the system of assessments used to make competency determinations differs between schools and classrooms, WPS uses an interim assessment program, the Scantron Performance series, as a common assessment to compare the performance across all schools in the fall, winter, and spring.

Under the competency-based education model, variable timing of assessments is a key concept. That is, because individual students are expected to have varying trajectories and rates of learning, student mastery would be evaluated at different time points throughout the year. To align with the fundamental concept of treating learning as variable across students, WPS would prefer to administer grade level state summative assessment at different times throughout the year for different students, based on when students have demonstrated readiness to move to the next level or grade. As noted by Dr. Grenham, WPS sees value in the rigor represented by the current state summative assessment (the CMAS Math and English Language Arts assessments), and the district is not seeking to eliminate use of the CMAS tests. WPS is instead seeking flexibility from the state to determine when individual students take these assessments.

WPS uses an approach similar to New Hampshire's PACE project by working with teams of teachers during the school year to ensure teachers understand the expectations and skills required to move students toward competencies. While the WPS system is not performance-based, district curriculum and instructional staff work with teachers to provide them with an understanding of the type of tasks or assessments needed to elicit evidence that students have mastered a given competency. The district also helps teachers clearly communicate expectations and criteria for success to students. WPS staff claim the system is still in its "infancy," and they expect to spend several years building teacher capacity to implement the proposed system. This includes calibrating expectations around success criteria for each content area and helping teachers improve their use of assessments to evaluate student progress toward meeting the desired competencies.

Figure 5 in Appendix A presents information on costs associated with investing in the Scantron Performance Series, reporting the data, and convening professional development sessions with teachers.



ASSESSMENT QUALITY CRITERIA

According to current ESSA regulations, innovative assessment systems used under the demonstration pilot must:

§1204(e)(2)(A)(ii). Be aligned to the challenging state academic standards and address the depth and breadth of such standards.

§1204(e)(2)(A)(iv). Generate results that are valid and reliable, and comparable, for all students and for each subgroup of students...as compared to the results for such students on the State assessments.

§1204(f)(1)(B)(i). [Be] comparable to the State assessments under section 1111(b)(2)(B)(v), valid, reliable, of high technical quality, and consistent with relevant, nationally recognized professional and technical standards.

Most of the concepts that we describe, and most of those discussed in the educational measurement field (including USED's peer review criteria), pertain primarily to the evaluation of single tests or assessments. Although the demonstration pilot allows states to pilot innovative assessment systems, there is little clarity about exactly what constitutes such a system (as opposed to an individual assessment). Our discussion thus focuses on criteria that could be used to evaluate individual tests or assessments used within such a system. As noted above, a fundamental requirement of the demonstration pilot is that any assessment system adopted by the state must produce annual achievement determinations for each student. Many of the criteria and concepts we describe below can also be applied to these annual achievement determinations.

In what follows, we first discuss some general design criteria for assessment systems. Then, we provide a more detailed discussion of key assessment criteria mentioned explicitly in federal and CO legislation, describing how these relate to the demonstration pilot. We then consider how these criteria are relevant for the assessment system examples above. Lastly, we describe two nationally recognized assessment criteria frameworks, one from the Council of Chief State School Officers (CCSSO) and one from the USED Peer Review system. These frameworks are included to illustrate the breadth and depth of evidence needed to fully evaluate the quality of assessments used in an accountability system, which necessarily goes beyond the concepts described here. Note that we are not providing a complete evaluation of the assessment system examples described above. Rather, we are illustrating some of the key issues that would need to be addressed in such an evaluation. This section is not intended to provide the state or local districts with guidance on how different assessment systems should be modified to meet quality criteria or the requirements of the demonstration pilot.



ASSESSMENT SYSTEMS

The National Research Council's "Knowing What Students Know: The Science and Design of Educational Assessment" (Pellegrino, Chudowsky, & Glaser, 2001), a seminal report on educational assessment, articulated a number of important concepts to guide the design of assessment systems intended to promote student learning. First, the report emphasizes that "one form of assessment does not serve all purposes" (p. 252) and that, "it is inevitable that multiple assessments (or assessments consisting of multiple components) are required to serve the varying educational assessment needs of different audiences" (p. 252). The report envisioned designing coordinated systems of assessment that are comprehensive, coherent, and continuous. A comprehensive system is one that draws on a range of measurement and assessment approaches to support educational decision-making. This could include a combination of classroom-based assessment activities as well as statewide standardized forms of assessment. A coherent system is one in which the "conceptual base or models of student learning underlying the various external and classroom assessments within a system should be compatible" (p. 255). The report describes both horizontal and vertical dimensions of coherence. Vertical coherence refers to a system in which assessments used at different levels – classroom, district, state – are all consistently tied to the same learning goals. Horizontal coherence implies that curriculum, instruction, and assessment are well integrated with one another at each level of the system. This entails more than simply ensuring that all assessments are "aligned" to the same content standards. Finally, a continuous assessment system is one that provides information about student learning over time, not only snapshot measures of achievement at single points in time.

We emphasize three key takeaways from these recommendations. First, the design of an assessment system must begin by stating the intended uses of the assessment system. These uses could be summative (e.g., making annual achievement determinations for each student) or formative (e.g., providing information to teachers that can assist them in planning subsequent instruction). We note that the demonstration pilot legislation requires innovative assessment systems to fulfill both of these purposes, in addition to others. Second, the system should include different, yet coordinated, components that are ideally suited to each of these purposes. In a recent article about the design of assessment systems, Shepard, Penuel, and Davidson (2017) write that, "creating a coherent and effective assessment system between classroom and statehouse does not mean building a *single* instrument to serve both formative and summative purposes" (p. 52). Third, there should be a clear model of learning and evidence base used to develop the assessment system. Penuel and Shepard (2016), for example, describe four potential models of formative assessment and consider the implied models of student learning underlying each one. These types of research-based syntheses can provide valuable information when developing different parts of the assessment system. The criteria we outline below should also be considered during the development and evaluation of an assessment system.



DEFINITIONS OF QUALITY CRITERIA

The *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), jointly published by the American Educational Research Association, American Psychological Association and National Council on Measurement in Education, provides the consensus view of experts in the field of testing and assessment regarding the criteria for developing, using, and evaluating tests. The *Standards* were used in the development of both the CCSSO and USED assessment criteria, and should inform the evaluation of any assessment or assessment system.

According to the *Standards*, validity is the most fundamental concern when developing or evaluating tests and assessments. The *Standards* define validity as, “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA & NCME, 2014, p. 11). This definition highlights three important points when considering whether a test is “valid.” First, validity refers to the quality of a test for a particular use, not to a test itself. A test that is valid for one use may not be valid for a different use. Second, an evaluation of the validity of a test requires drawing on a diverse array of theoretical and empirical evidence. The *Standards* outline five primary sources of validity evidence that are often used to evaluate the validity of proposed inferences.⁴ Third, judgments about validity are not “all or none,” but rather indicate the degree of support for proposed interpretations and uses.

This concept of validity highlights the importance of clearly articulating the intended uses and interpretations of an assessment. The demonstration authority legislation also requires stating the rationale for implementing an innovative assessment system. Clearly stating and evaluating the intended aims and uses of the assessments is thus required both by professional standards and legislation. If assessment results will be used for multiple purposes within the system, each of these uses needs to be clearly stated and evaluated. If the assessment system will utilize pre-existing tests developed in other contexts, it is still necessary to evaluate whether these tests are also valid for their proposed use in the current assessment system.

The argument-based approach to validation (Kane, 2006; 2013; 2016) is one current framework for evaluating tests or assessments with attention to intended uses. According to Kane:

“An argument-based approach to validation involves two basic steps: (a) specify the claims that are to be based on test scores, as an interpretation/use argument, and (b) evaluate the plausibility of these claims using appropriate methods and evidence in a *validity argument*” (2016, p. 309)



⁴ The five sources of validity evidence described in the *Standards* are: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on the consequences of test use.

We view the demonstration pilot’s statements about reliability, comparability, and alignment as a requirement to evaluate, in particular, whether claims that results of tests in the system are reliable, comparable, and aligned to the state standards are warranted. This is particularly important for the annual achievement determinations. In Table 2, we consider what claims about reliability, comparability and alignment might look like in the context of the demonstration pilot and how they might be evaluated. We consider three ways assessment results might be used under the pilot: 1) determination of year-end achievement status for individual students, 2) determination of year-end school ratings for accountability, and, 3) providing useful feedback to teachers and students that can inform classroom instruction.

Table 2. Important assessment characteristics to evaluate.

Characteristic	Intended Use of Assessment		
	Student Year-End Achievement Determinations	School Accountability Determinations	Inform Classroom Instruction
Reliability the consistency and generalizability of test scores	<ul style="list-style-type: none"> Ensure that student year-end proficiency designations are consistent. 	<ul style="list-style-type: none"> Evaluate the reliability of <u>aggregate</u> school-level scores, which may differ from student scores. 	<ul style="list-style-type: none"> Consistency of scores may be helpful, but teachers and students may factor in additional context outside of the scoring guide to inform ratings. Transparency in scoring rules may be more important than achieving consistency across students.
Comparability the degree to which test scores/results can be directly compared	<ul style="list-style-type: none"> Ensure that a student’s achievement designation obtained in one district would be the same in a different district, especially if determinations are based on a mix of evidence that varies across districts. 	<ul style="list-style-type: none"> Ensure that a school’s test-based accountability determination obtained in one district would be the same in a different district, especially if based on a mix of evidence that varies across districts. 	<ul style="list-style-type: none"> Comparability would be important to achieve within classrooms (across students in a classroom or over time during the year) but less critical across schools from a classroom teacher perspective.
Alignment the extent to which the content and cognitive demands of a test are consistent with those in a set of content standards	<ul style="list-style-type: none"> Ensure alignment of the content/cognitive skills covered in the body of assessment evidence to the state standards. 	<ul style="list-style-type: none"> Ensure alignment of the content/cognitive skills covered in the body of assessment evidence to the state standards. 	<ul style="list-style-type: none"> Ensure alignment to structure of local (district) instructional goals and tasks.

As indicated by Table 2, three key criteria that would be examined under HB 16-1234 and the demonstration pilot apply differently to assessments used for different purposes. This illustrates a key challenge faced by assessment systems that feature a single assessment intended to serve many different purposes.



Namely, when an assessment is used in a system for high-stakes school accountability purposes, it may need to meet design constraints that undermine its use for other purposes, such as informing instruction.

We next define each of the three criteria (reliability, comparability and alignment) explicitly referenced in HB 16-1234 and the demonstration pilot in more depth. We then briefly discuss the issue of test fairness.

Reliability

Reliability refers to the consistency and generalizability of test scores. At an initial level, we assume that scores are consistent across occasions. For example, if a student takes a test in the morning versus the afternoon, or had taken a test tomorrow instead of today, we assume they would get similar scores. Similarly, if a test or assessment is constructed by sampling tasks (e.g., items or prompts) from a larger domain of possible tasks, we assume the score a student receives on one set of tasks is a good indicator of the score they would have received on another, equivalent, set of tasks. When a test or assessment task requires scoring by a trained rater, we assume that there is a consistent scoring process and students would have received the same score if a different, equally qualified rater had been assigned to score their work. Reliability of scores will always be a matter of degree, rather than an all or none judgment.

The importance of reliability is related to how an assessment is used. Score reliability is particularly important when scores are used for high-stakes decisions (e.g., in an accountability system), but reliability may be less important if scores are used for lower-stakes purposes such as on-going classroom assessment to inform instruction. It is also critical to attend to the reliability of derived or aggregated scores, especially when these derived scores are used for high-stakes accountability purposes (Hill & DePascale, 2003; Haertel & Ho, 2016). When a standard setting procedure is used to set cut scores to determine student proficiency levels, for example, the consistency of the proficiency level classifications needs to be evaluated (Livingston & Lewis, 1995). When scores are aggregated to the teacher or school level, the reliability of the aggregate scores needs to be evaluated. The statistical properties of derived or aggregated scores may not always mirror those of the individual student scores.

In the context of HB 16-1234 and the demonstration pilot, the primary concern is with the reliability or consistency of the year-end achievement determinations for each student. However, because these determinations are also used to support inferences about aggregated school-level results, evaluating the reliability of these aggregate data is also relevant.



Comparability

Score comparability refers to the degree to which test scores or inferences about achievement for different students can be directly compared. If two students take identical tests, there is often a strong warrant for making the claim that scores are comparable. When two students take different tests, or different versions of the same test, it is more difficult to determine whether scores across the tests can be meaningfully compared. When tests are deemed comparable, we infer that regardless of which test a student takes, we would reach the same conclusion about the student's level of achievement. As with reliability and validity, comparability is also a matter of degree.

There are many reasons scores from different tests may not be comparable. As one example, the actual score scales for the tests may differ. Consider comparing scores on Test A with scores on Test B. Scores on Test A range from 50 to 100, while scores on Test B range from 1 to 10. Clearly, a direct comparison of scores across tests would be problematic. Suppose the score scales were adjusted to have similar values, or were used to make determinations about which proficiency level a student's score represents. The values or labels of the test results would now look similar, but still may not be comparable in a meaningful sense. When tests are written by different publishers or are constructed from different blueprints, the two tests may not measure exactly the same construct. The two tests would thus not be expected to lead to the same inferences for all students and the results would not be considered comparable.

Under the demonstration pilot, there are two especially relevant dimensions of comparability. First, achievement determinations must be comparable for all students and schools participating in the pilot. For systems that use the same set of tests or assessments across all participating schools, this may be relatively straightforward to document. In systems that allow participating schools or districts to select their own assessments, however, documenting comparability could prove much more challenging. Second, year-end achievement determinations based on the innovative assessment system must be comparable to determinations made using the existing statewide assessment, because during the pilot period some districts may continue to use the existing state assessment. This pertains to both student-level achievement determinations and school-level results.

Briggs and D'Agostino (2016)⁵ provide a detailed description of psychometric criteria for evaluating score comparability and outline the challenges that Arizona's menu of assessments plan is likely to face when attempting to verify comparability. Evans and Lyons (2017) discuss challenges associated with evaluating



⁵ The full text of this memo, prepared by Derek Briggs and Jerry D'Agostino for the Arizona Department of Education, is located in Appendix F.

comparability for the New Hampshire PACE system as well as methodologies that could be used to address some, but not all, of these challenges. A recent report submitted to USED by the National Center for the Improvement of Educational Assessment (2016) describes a number of different methodologies that could be used to investigate score comparability for innovative assessment systems used under the demonstration pilot. There remains a lack of consensus regarding the best methods for evaluating comparability under the demonstration pilot or about the degree of comparability such systems should be required to demonstrate.

Alignment

Alignment refers to the “extent to which the content and cognitive demands of an assessment tool are consistent with (or match) those given in a set of content standards or benchmarks that describe the curriculum with which the assessment was designed to be used” (National Council on Measurement in Education, n.d.). Alignment is a relevant concern for assessment systems that rely on new test development and for systems that use pre-existing tests.

Evaluating alignment is an inherently judgmental process. Most current methods for judging alignment stem from a report by Webb (1997), including a widely utilized and simplified approach described elsewhere (e.g., Webb, 2007). Judging alignment of an assessment to a set of standards under this method proceeds in two general steps. In the first step, the standards measured by the assessment are broken into pieces and each piece is categorized by the “depth of knowledge” (DOK) or level of cognitive complexity that meeting that piece of the standard requires. In the second stage, each item or task on the assessment is mapped back to one or more of the content pieces in the standards and also assigned a DOK level. Alignment is then determined based upon how well the assessment items or tasks cover the range of the standards and match the DOK levels of the standards. As Webb (2007) points out, this is a necessarily subjective process both in terms of matching test items to standards (or DOK levels) and in terms of determining how much coverage or alignment is sufficient to indicate “good” alignment. In addition, Webb (2007) notes that, “The results produced from the [alignment study] pertain only to the issue of agreement between the state standards and the assessment instruments...the alignment analysis does not serve as external verification of the general quality of a state’s standards or assessments” (p. 9).

When evaluating alignment (and other criteria described above), two threats to validity are “construct underrepresentation” and “construct irrelevant variance.” Construct underrepresentation occurs when important aspects of the construct intended to be measured are not adequately represented by the test. Underrepresentation can occur when there are insufficient items on a test evaluating a given standard, or when the cognitive complexity required to answer test items is lower than the complexity of skill described in the standards.



Construct irrelevant variance occurs when factors not intended to be measured by the test systematically affect performance. For example, if math questions in a given assessment are framed as word problems, a student with lower reading comprehension skills may not perform as well as other students with higher reading comprehension skills.

In the context of HB 16-1234 and the demonstration pilot, attention needs to be given to the alignment of tests or assessments to both the state standards and to the particular scope and sequence of each district's curriculum. For the year-end student achievement determinations, alignment of assessment content to the state standards is critically important. But if the assessment system is also intended to improve instruction, alignment between the assessments and the instructional setting in schools and classrooms also needs to be considered. Forte (2016) provides a more extended discussion of alignment, including alternative methods for evaluating alignment, and discusses these issues in the context of the USED Federal Peer Review criteria.

Fairness

The *Standards* include an entire chapter discussing fairness, emphasizing the critical importance of evaluating the fairness of tests and assessments. The *Standards* also acknowledge that, "the term *fairness* has no single technical meaning and is used in many different ways in public discourse...individuals [might] endorse fairness in testing as a desirable social goal, yet reach quite different conclusions about the fairness of a given testing program" (p. 49).

Nonetheless, the *Standards* describe four views of fairness to be considered: 1) equitable treatment of all test takers during the testing process, 2) the absence of measurement bias, 3) equitable access to the constructs measured, and 4) validity of individual test score interpretations. Ensuring score comparability as described above is one aspect of ensuring fairness implied by these views. Another critical aspect entailed by these views is ensuring accessibility of the tests or assessments for all students. For example, if modifications or adaptations are not made to the assessments or tasks used in the innovative system to meet the particular needs of students with certain disabilities, or of students who are English language learners, then the resulting scores could be biased. That is, the scores reported from assessments that are not accessible to certain groups of students will not accurately reflect what those students know and can do. Another fairness consideration pertains to whether schools and teachers serving students from different populations are equally well qualified and prepared to implement the innovative assessment system.

Although "fairness" is not called out directly in the demonstration pilot language or in the HB 16-1234 language, it remains a critical aspect to consider when evaluating assessments and assessment systems. The demonstration pilot legislation does

require, for example, that the innovative assessment system be accessible for all students; that stakeholders representing the interests of children with disabilities, English learners, and other vulnerable children be consulted in the design process; and that states provide evidence documenting that all participating schools have the capacity to implement the innovative assessment system. Examining fairness with regard to test design and reporting will be at the forefront of any quality review of assessments used to make annual proficiency determinations. Peer reviewers and CDE staff evaluating the individual assessments and tasks used in a proposed innovative assessment system will have to check that these are designed to address fairness and accessibility for all students. A recent edited volume published by the National Council on Measurement in Education (NCME), describes many of the technical methods used to evaluate fairness in educational assessment (Dorans & Cook, 2016).

Additional Criteria

We focused on reliability, comparability, alignment, and fairness because these concepts are the ones referenced in the demonstration authority pilot and because they are critical for evaluating any assessment. There are many other factors that contribute to determining the quality of an assessment or assessment system. These include test security and quality control, test development procedures, consideration of administrative support for test administration, scoring and reporting, formatting score reports to facilitate appropriate interpretations, and many others. It is beyond the scope of this document to discuss these additional criteria at length. As mentioned above, assessments need to be evaluated in regards to their intended use. This should also include attention to potential unintended consequences of implementing the assessment system, particularly in regards to fairness and equity concerns. Many of these additional criteria are described in the national assessment criteria frameworks we list below.

CHALLENGES FOR MEETING QUALITY CRITERIA IN INNOVATIVE ASSESSMENT SYSTEM EXAMPLES

Having discussed key criteria that would be considered under HB 16-1234 and the demonstration pilot, we turn to considering validity, reliability, comparability, and alignment in the context of the assessment systems described above. In an assessment system that bases year-end achievement status on multiple assessments administered during the school year, it may not be realistic for every assessment in the system to fully satisfy all criteria. For example, in New Hampshire's PACE project, not every performance task will be aligned to the entire state standards, a point emphasized by the New Hampshire Department of Education. The relevant question is whether the body of evidence represented by the performance tasks used to reach the final achievement determination meets these criteria. The exact criteria each assessment needs to meet will depend upon how it is used within the system.



Selecting and implementing an assessment system involves tradeoffs, both in terms of the concepts described above, and in other areas such as cost. These tradeoffs can be seen clearly in the cases of New Hampshire and Arizona. In the PACE project, the performance assessments may enhance the validity of score inferences by requiring students to demonstrate more complex and authentic skills. In addition, the participation of educators in the assessment development and implementation process provides an opportunity to improve instruction and classroom assessment practices. The teacher-developed assessments are also more likely to be well integrated with the local instructional context. On the other hand, it is difficult to cover as much breadth of content with performance assessment tasks, each of which requires more time to complete. As a result, reaching full content alignment with a state's content standards may be difficult. Moreover, the level of reliability and comparability that can be achieved across performance tasks is generally not as high as can be achieved with more standardized multiple-choice tests (Haertel & Linn, 1996). It is unclear whether the PACE districts will be able to satisfy the comparability requirements of the demonstration pilot at this point. Lane and DePascale (2016) discuss these and other tradeoffs involved when using performance assessments in accountability systems. Finally, implementing a system such as the PACE project requires extensive resources and supports, and is time-consuming, both for the state administrators and educators (Evans & Lyons, 2017). Even though the initial per pupil cost estimates (see Appendix A) should decrease in the long-run, this assessment system requires convening large groups of teachers each year, both within and across districts, to continuously calibrate expectations based on the different assessments used to make annual competency determinations.

A system using pre-existing, external assessments, such as Arizona's, is likely to be much less expensive for the state. Although these assessments may be logistically easier to administer in each district, the administration and reporting process may not be easy to coordinate across districts using different assessments. Allowing districts the freedom to select their own assessments may increase the (perceived) relevance of the tests for classroom instruction, but since these tests are externally developed and often intended to be curriculum-neutral they may not be as closely related to instruction as assessments in the PACE project. Scores from these tests, which usually rely primarily on multiple-choice items, also tend to be quite reliable. On the other hand, the multiple-choice items on these tests may not provide as much evidence about students' higher-order cognitive skills (such as those included in the Colorado Academic Standards) relative to more complex performance tasks. Moreover, Briggs and D'Agostino (2016) describe why it will be challenging, if even possible, to provide strong evidence of comparability across districts when each district selects a different test. In brief, it is unclear how a state would be able to go about documenting comparability based on assessment results from numerous different vendors. Documenting comparability under this system would either

require students to be double-tested (by using a common test across districts in addition to the district-selected tests) or that common items be embedded in the tests (which could be difficult or impossible for tests produced by different vendors). These methods of ensuring comparability would undermine the goal of reducing testing time and would be extremely difficult to coordinate with multiple test vendors. It is also unclear how well-aligned these tests would be to any specific state's standards. Hence, although pre-existing tests may have higher levels of reliability, and could improve the reliability of achievement determinations, such a system will face challenges meeting comparability and alignment requirements.

The local Colorado systems described above are likely to face a mix of the challenges discussed for systems like those in New Hampshire and Arizona. Ensuring alignment and comparability will be a challenge for the S-CAP districts, similar to the challenges described for the two state models. Although Westminster Public Schools plans to use a common interim assessment that may help support comparability claims, it would be highly unlikely that the multiple-choice Scantron Assessment currently meets the level of rigor and alignment to state standards represented by the CMAS assessments. Some may question whether Scantron serves as the right benchmark for calibrating performance expectations across all other assessments used by the schools. The Cherry Creek system, which uses a single year-end summative test with integrated interim assessments, appears to represent the most balanced assessment system. This system would likely not face the same challenges when documenting comparability and reliability, although determining alignment to the state standards may be a challenge. However, we note that this system appears to represent primarily a substitution of one large-scale standardized assessment (ASPIRE) for another (CMAS), rather than an innovative system of student assessment. Moreover, the system does not reduce testing time, which is one of the key stated goals in HB 16-1234. Districts that believe the ACT suite of assessments are directly aligned with college readiness goals may embrace these assessments as a substitute for the current system. However, for those districts that place a high value on locally developed or selected assessments, such as the teacher-developed assessments in Buena Vista or the teacher-selected assessments in Westminster Public Schools, adopting the Cherry Creek system would likely raise concerns. It seems unlikely that all districts in Colorado would want to give up their own locally-selected assessments and use a state-mandated interim-based system.

These challenges highlight an inherent tension in the demonstration pilot requirements: the proposed assessment systems aim to be innovative but also need to produce the same achievement results that would have been obtained with the current assessments. In addition, the demonstration pilot legislation requires the innovative assessment system to serve a number of distinct purposes that are unlikely to be well-served by any single form of assessment.

EXISTING ASSESSMENT QUALITY FRAMEWORKS

As indicated above, a thorough evaluation of all assessments used to make annual proficiency determinations for students requires many additional considerations beyond those described above. To provide a sense for the scope and nature of these considerations, we identify two comprehensive sets of criteria that could serve as useful resources when developing and evaluating assessments and assessment systems. The first is the Council of Chief State School Officers' (CCSSO, 2014) "Criteria for Procuring and Evaluating High-Quality Assessments," which provide criteria for evaluating large-scale summative assessments. The second is the USED (2015) "Critical Elements for the State Assessment System Peer Review," which guides the peer review of state assessment systems.

CCSSO Criteria

The CCSSO document describes criteria for evaluating summative assessments intended to measure college and career readiness. This document lists 10 general criteria, as well as 9 criteria specific to evaluating the alignment of tests to state standards in English Language Arts/Literacy and 5 criteria specific to evaluating the alignment of tests to state standards in Mathematics. The document also describes potential state-specific criteria that might be considered. The full list of criteria is included in Appendix C; the criteria include references to the concepts of validity, reliability, comparability, alignment, and fairness discussed above.

The National Center for the Improvement of Educational Assessment (2016a, 2016b) has produced an extensive set of resources to support organizations evaluating particular tests or assessments relative to the CCSSO criteria. These supporting documents re-organize the criteria into two primary categories – test content and test characteristics. The evaluation framework described includes: articulation of the claims that need to be evaluated, description of what constitutes sufficient evidence to support the claims, examples that illustrate the evaluation process, and a summary of key connections to the *Standards*.

Federal Peer Review Critical Elements

The USED (2015) has provided a set of non-regulatory "Critical Elements" to be considered in the peer review of state assessments. The map of all critical elements is included in Appendix D. This is an extensive set of criteria that expert peer reviewers use to evaluate current state assessment systems used for accountability. The demonstration pilot calls for a peer review process to evaluate innovative assessment systems (§1204 (f)), and it seems plausible that the criteria required for the demonstration pilot peer review will be similar to those described in the USED documents. The extensive nature of the critical elements underscores the magnitude of conducting a comprehensive evaluation of either new or existing assessment systems.



The federal peer review guidelines also address reliability, comparability, alignment, and fairness. As examples, Critical Element 4.1 of the federal criteria addresses reliability; Critical Elements 4.2 and 5.1-5.4 address issues related to comparability and fairness; and Critical Element 3.1 addresses alignment. Examples of the evidence required from states for each critical element that must be addressed for peer review are located in Appendix E.



CONCLUSION

We conclude this report by highlighting three points. First, we remind readers that none of the assessment systems reviewed above is fully implemented, and some are not yet fully designed. Each of the examples would likely need additional development or evidence in order to fully meet the criteria outlined in the demonstration pilot or the CCSSO and Federal Peer Review Guidelines. Second, state-specific legislation in Colorado regarding the measurement of student growth adds an additional set of requirements that will need to be considered when designing an assessment system for the demonstration pilot; we discuss this further below. Third, and most importantly, when developing an assessment system for the demonstration pilot it will be critical to explicitly state the primary goal or goals of the system. Careful attention should then be paid to how the proposed assessment system is designed to meet these goals and whether the system can realistically meet all the goals. We now discuss each of these points in more detail.

Because the five systems described above are still undergoing development and refinement, none of the systems appears ready to *fully* satisfy the demonstration pilot criteria or the CCSSO and USED criteria reviewed above. Each of the systems we discussed is likely to face different challenges moving forward. Ensuring comparability of assessment results across districts and with the existing state assessment will likely be the most difficult challenge, and will be on-going. Although New Hampshire's recent report to USED indicates, for example, that they believe the PACE project demonstrates sufficient levels of comparability, this evidence is based on the results from a limited set of districts (i.e., the initial eight pilot districts) selected to participate in the pilot due to demonstrated "readiness" to implement the requirements of this work (NHDOE, 2016). It remains to be seen if such an effort can be sustained over time for the 141 districts that have not yet participated in the PACE pilot. Designing an innovative assessment system that can continuously meet all of the demonstration pilot requirements is a lengthy and expensive undertaking.

Colorado faces the additional concern that any proposed assessment system must also fulfill state legislation regarding the measurement of student growth. Specifically, SB 09-163 requires the use of the "Colorado Growth Model" (i.e., student growth percentiles) as a "common measure to describe how much academic growth each student needs to make" (22-11-102 (2)(a)). Some of the assessments used in the S-CAP districts (e.g., local assessments created by teachers in Buena Vista) may not have the technical properties required to construct growth percentiles. Although New Hampshire uses and reports student growth percentiles for school accountability, the PACE project will not and cannot use growth percentiles. In addition, although growth percentiles could in theory be constructed at the district level for some of the assessments mentioned above

(e.g., the NWEA MAP or ACT ASPIRE tests), the norm groups or academic peer groups used to construct the growth percentiles would not be reflective of the entire state, and therefore the interpretations of these growth percentiles would not hold across different districts. Many district student populations are also too small for constructing district-based growth percentiles. Other approaches to evaluating growth such as “value-tables” (Castellano and Ho, 2013) could be employed, but using these alternative methods would require an amendment to existing Colorado statute.

The design and evaluation of an assessment system used for the demonstration pilot needs to be guided by clear statements of the purpose of such a system, including how the different components will meet those purposes. The requirements built into the demonstration pilot legislation may be at odds with some of these goals. As an example, we consider two potential goals the state may have in designing an innovative assessment system for the demonstration pilot: providing more instructionally useful information to teachers and students (i.e., serving formative assessment purposes) and reducing overall state-mandated testing time. However, the resulting system must also produce highly reliable, valid, and comparable achievement determinations relative to the state content standards for every student in each of the federally required grades and subjects. The most efficient way to accomplish this latter aim may be with a statewide year-end test similar to what most states (including Colorado) currently use. None of the examples we discussed above appear to have settled upon a system design that adequately accomplishes this requirement without reliance on a statewide assessment. Moreover, it is unclear whether the models described above will necessarily accomplish either of the other two aims. The models used in Arizona and New Hampshire could actually be seen as increasing the amount of state-mandated testing – by replacing the use of a single year-end assessment with performance assessments or interim assessments administered multiple times per year to each student. While the hope is that such assessments will better support teaching and learning, this is not guaranteed.

The requirement to continue making comparable annual achievement determinations for all students makes the design of truly innovative systems under the demonstration pilot difficult. In an attempt to reduce testing time, one option is to utilize the same assessments for both formative and summative accountability purposes. Yet as noted by Shepard et al. (2017), “Painful lessons from the past... remind us that creating a coherent and effective assessment system does not mean building one assessment instrument to serve both formative and accountability purposes” (p. 52). In the past, efforts to design large-scale accountability tests that were also “tests worth teaching to” were undermined by requirements that these tests be affordable and highly standardized, which made them “ill-suited to serve as models for high-quality teaching and learning at the local level” (p. 52). The other alternative is to design a coherent system that uses

multiple different forms of assessments for these varying purposes – large-scale, external assessments for summative accountability decisions and locally developed forms of assessment to improve and inform instruction. This approach, however, could be viewed as increasing the amount of state-mandated testing. This may be the case even if, as is recommended by experts (e.g., Pellegrino et al., 2001), the assessments used to inform and improve instruction are locally developed and well-integrated into the local school or district curricula. A practical challenge of this latter approach is also the need to devote resources to the training needed to implement innovative assessment practices in schools.

Haertel (2009) describes a number of innovative assessment practices that could be used to improve accountability testing, including the use of matrix sampling, performance assessments, and school-based portfolios. Many of these would require changes to current accountability rules, including “decoupling the multiple purposes for which some tests are used” (p. 9) and relaxing requirements to test every student in every subject or produce state-mandated growth metrics. These approaches also raise new challenges; matrix sampling, for example, may be difficult or not feasible to carry out in small rural districts where in some locations an entire district may enroll fewer than 50 students. In these cases, different approaches to evaluating school performance may need to be explored, although this is also likely to be true with more standard test administration procedures.

These tradeoffs highlight a fundamental tension embedded in the demonstration pilot requirements. Specifically, the demonstration pilot requirement to make year-end achievement determinations that are comparable to determinations based on existing state tests, for all students in all tested subjects, severely limits the nature of assessment systems that can be pursued under this pilot. Coupled with existing Colorado statutes regarding the measurement of student growth, it is unclear how the systems reviewed above could satisfy all of these requirements simultaneously in an effective manner. Truly innovative systems of assessment may not conform to existing accountability rules, either in federal or state legislation, and hence may not be possible under the demonstration pilot at the present time. A critical step at this point is to prioritize the goals of adopting an innovative assessment system and identifying which (if any) of the state legislative requirements may be up for revision. These priorities and constraints can then guide the design and evaluation of potential assessment systems.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Briggs, D., & D'Agostino, J. (2016). A Technical Commentary on Arizona's Menu of Assessments Legislation (H.B. 2544) [White Paper].
- Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers. Retrieved from: <http://www.ccsso.org/Documents/2013GrowthModels.pdf>
- Council of Chief State School Officers. (2014). Criteria for Procuring and Evaluating High-Quality Assessments. Retrieved from: <http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf>
- Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. New York, NY: Routledge.
- ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016). Retrieved from: <https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf>
- Evans, C., & Lyons, S. (2017). Comparability in innovative assessment systems for state accountability. *Educational measurement: Issues and practice*. doi:10.1111/emip.12152
- Forte, E. (2016). Evaluating Alignment in Large-Scale Standards-Based Assessment Systems. Washington, DC: Council of Chief State School Officers. Retrieved from: <http://www.ccsso.org/Documents/TILSA%20Evaluating%20Alignment%20in%20Large-Scale%20Standards-Based%20Assessment%20Systems%20-%20FINAL.pdf>
- Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington, D.C.: National Center for Education Statistics.
- Haertel, E. H. (2009). *Reflections on educational testing: Problems and opportunities*. New York, NY: Carnegie Corporation of New York-Institute for Advanced Study Commission on Mathematics and Science Education. Retrieved from <http://www.csai-online.com/sites/default/files/resource/imported/b9ca12a8-9d04-404d-87ae-1e0013ff1bcb.pdf>
- Haertel, E., & Ho, A. (2016). Fairness using Derived Test Scores. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 217-237). New York, NY: Routledge.

- Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice*, 22(3), 12–20. <https://doi.org/10.1111/j.1745-3992.2003.tb00133.x>
- House Bill 16-1234 (2016). Retrieved from: https://leg.colorado.gov/sites/default/files/documents/2016a/bills/2016A_1234_enr.pdf
- House Bill 2016-2544. (2016). Retrieved from: <http://www.azleg.gov/legtext/52leg/2r/bills/hb2544p.pdf>
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice*, 23(2), 309–311. <https://doi.org/10.1080/0969594X.2016.1156645>
- Lane, S., & DePascale, C. (2016). Psychometric considerations for performance-based assessments and student learning objectives. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 77–106). New York: Routledge.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32(2), 179–197. <https://doi.org/10.1111/j.1745-3984.1995.tb00462.x>
- National Center for the Improvement of Educational Assessment. (2016a). CCSSO Criteria for High Quality Assessments: Focus on Test Content. Retrieved from: http://www.nciea.org/sites/default/files/publications/Guide-to-Evaluating-CCSSO-Criteria-Test-Content_020316.pdf
- National Center for the Improvement of Educational Assessment. (2016b). A Guide to Evaluating College- and Career-Ready Assessments: Focus on Test Characteristics. Retrieved from: http://www.nciea.org/sites/default/files/publications/CFA-TestCharacMethod-EvalMethod_Final.pdf
- National Council on Measurement in Education. (n.d.) Glossary of Important Assessment and Measurement Terms. Retrieved from: https://www.ncme.org/ncme/NCME/Resource_Center/NCME/Resource_Center/Glossary1.aspx?hkey=8bd573bd-a7b4-498a-93b9-3e0081c557c0
- New Hampshire Department of Education. (2015). *New Hampshire Performance Assessment of Competency Education: Progress Report to the United States Department of Education*.

- Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., 787-850). Washington, DC: American Educational Research Association.
- Senate Bill 09-163. (2009). Retrieved from:
<https://www.cde.state.co.us/sites/default/files/documents/cdedepcom/download/pdf/senatebill163.pdf>
- Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, 19(1), 405-450. <https://doi.org/10.3102/0091732X019001405>
- Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan*, 98(6), 47-52. <https://doi.org/10.1177/0031721717696478>
- U.S. Department of Education. (2015). U.S. Department of Education Peer Review of State Assessment Systems: Non-Regulatory Guidance for States for Meeting Requirements of the Elementary and Secondary Education Act of 1965, as Amended. Washington, DC: Office of Elementary and Secondary Education. Retrieved from: <https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf>
- Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Research Monograph No. 6). Washington, DC: Council of Chief State School Officers. Retrieved from: <http://files.eric.ed.gov/fulltext/ED414305.pdf>
- Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7-25. <https://doi.org/10.1080/08957340709336728>

APPENDIX A: COST ESTIMATES FOR ALTERNATIVE ASSESSMENT MODELS

Note: The cost estimates provided for the SCAP districts reflect costs for developing a common framework for their accountability system since different SCAP districts intend to use different assessments for accountability purposes. In the case of Arizona, since the costs of implementing the assessment system proposed will be borne largely by the districts, no information on costs can be provided by the state. However, for Arizona and for some of the SCAP districts, interim assessment costs for the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessments shared by Cherry Creek from the 2014-15 school year serve as a good proxy for estimating per pupil costs for districts using MAPs in Arizona and in the SCAP. Cost estimates presented in this report should be treated with caution since in all cases, these estimates do not encompass all direct and indirect costs associated with implementing and maintaining a given assessment system over time. Additionally the pupil cost estimates cannot be directly compared since the cost estimates shared are based on different assumptions.

Figure 1. New Hampshire per pupil cost for PACE Project in 2014-15.

Source: NH DOE

1. State personnel costs to support project - **\$250,000**
2. Technical consultant costs - **\$1,600,000**
3. Disbursements to districts - **\$300,000** (this includes stipends for summer work and cross district meeting costs, as well as grants to each district for their local costs. The grants only cover bargaining-agreement-required teacher stipends for out-of-contract work days, substitute teachers, mileage, and directly-related supplies.
4. Communications and PR - **\$150,000**

Estimated per pupil costs: \$169.17 (total enrolled in 8 districts = 13,596)

Note: Costs reflected are only from year one of this pilot and do not include costs for the local assessments used by each district used in addition to the common tasks. Per NH DOE, the technical consultant costs are also likely higher than the amount noted above and these costs will continue through the end of the pilot period in 2016-17.



Figure 2. Illustrative per pupil costs for Arizona districts using NWEA MAPs.
Source: CCSD

- > In Arizona's proposed model, since assessments costs will be borne largely by districts opting to participate in this model, no cost estimates are available from the state.
- > Since many school districts in Arizona use NWEA MAPs, the below per pupil cost estimates are provided here for illustrative purposes from a school district (Cherry Creek) in CO who used NWEA MAPs during the 2014-15 school year.
 - Math, Reading, and Language license - **\$8.00** per student
 - Science license - **\$2.50** per student

Total estimated per student cost in a given district: \$10.50

Note: Estimated per student cost only covers contractual costs with vendor but does not include costs incurred by the school district such as test administration costs or professional development provided on data use. Per pupil costs come from the 2014-15 school year.

Figure 3. Costs for using ACT interim and summative assessments in Cherry Creek School District. Source: CCSD

For ACT Aspire:

- **\$25.00** per student for the summative assessment (Reading, English, Science, Math, optional Writing) but discounts applied if tests ordered in bulk. With discounts, total costs go down to **\$22.00** per student. The summative is administered once a year in either the Fall or Spring.
- The interim assessment is **\$9.00** per student but can be bundled with the summative and receive a **\$4.00** per student discount.

Estimated total cost per student for summative and interim (including discounts): \$26 per student.

Note: Costs reflected from 2016-17 school year reflect contract costs with ACT but does not include costs incurred by the district such as professional development costs, or test administration costs.

Figure 4. Costs for supporting the SCAP accountability development work.
Source: BVSD.

Estimated total cost associated with supporting the work to develop the accountability model as estimated by SCAP member districts:

- Estimated total for 2016-17 school year: **\$225,800**

Estimated per pupil cost: \$64.51 (assumes enrollment of 3,500 students across all SCAP member districts)

Note: Per pupil cost is not associated with any assessment related costs incurred but only reflect costs associated with training, developing and building out the new accountability system across SCAP member districts.

Figure 5. Costs for providing instructional and assessment supports for competency based education model in Westminster Public Schools. Source: WPS.

- > Scantron and DIBELS system – online assessments: **\$164,000**
- > Storing system for data: Alpine and Empower **\$61,000** for Alpine, **\$50,000** for Empower
- > Instructional pieces/professional development enacted to support use of the data: **\$240,000**

Estimated per pupil cost: \$29.90 (assumes enrollment of 10,000 students)

Note: Costs from 2015-16 school year do not include other optional assessments purchased by schools to evaluate students. Professional development costs will likely fluctuate over the years.

APPENDIX B: ASSESSMENT QUALITY REVIEW TOOL

Content Area: _____ Name of Assessment: _____
 Grade Level: _____ Date of Review: _____
 Reviewer (s): _____

PACE High Quality Assessment Review Tool

Part 1: Assessment Profile
<p>Items Submitted – check all that is submitted and <u>fully</u> completed:</p> <p><input type="checkbox"/> NH PACE Performance Task Template</p> <p><input type="checkbox"/> Teacher Instructions: materials needed, time required for administration, procedure</p> <p><input type="checkbox"/> Student Performance Tasks: what the student is required to do and produce (prompt, directions, materials, checklists, etc.)?</p> <p><input type="checkbox"/> Scoring Rubric</p> <p><input type="checkbox"/> Answer Key or Guidelines: <u>Please circle if Not Applicable</u></p> <p><input type="checkbox"/> Actual Texts or links to texts, videos, data charts, etc.</p>
<p>Performance Task Description:</p> <p><input type="checkbox"/> Fully describes the context, the anticipated activities, products and/or presentations, resources, texts, and materials needed, and what students are expected to demonstrate.</p> <p><input type="checkbox"/> Partially describes the context, the anticipated activities, products and/or presentations, resources, texts, and materials needed, and what students are expected to demonstrate.</p> <p><input type="checkbox"/> Minimally describes the context, the anticipated activities, products and/or presentations, resources, texts, and materials needed, and what students are expected to demonstrate.</p> <p>Comments:</p>
<p>Teacher Directions:</p> <p><input type="checkbox"/> Fully describes all aspects of the administration of the task including pre-requisite learning, lessons for scaffolding, what the students will do independently. These directions follow the guidance outlined in the document entitled “Guidelines for Independent Student Work Products for NH PACE Assessments: Implications for instructional scaffolding.”</p> <p><input type="checkbox"/> Partially describes the aspects of the administration of the task including pre-requisite learning, lessons for scaffolding, what the students will do independently. These directions partially follow the guidance outlined in the document entitled “Guidelines for Independent Student Work Products for NH PACE Assessments: Implications for instructional scaffolding.”</p> <p><input type="checkbox"/> Minimally describes aspects of the administration of the task including pre-requisite learning, lessons for scaffolding, what the students will do independently. These directions minimally follow the guidance outlined in the document entitled “Guidelines for Independent Student Work Products for NH PACE Assessments: Implications for instructional scaffolding.”</p> <p>Comments:</p>



Content Area: _____ Name of Assessment: _____
Grade Level: _____ Date of Review: _____
Reviewer (s): _____

PACE High Quality Assessment Review Tool

To what extent is scaffolding provided?

- No scaffolding is provided for aspects of the task that are being scored with the rubric
- Low level of scaffolding is provided for aspects of the task that are being scored with the rubric
- Some scaffolding is provided for aspects of the task that are being scored with the rubric
- High level of scaffolding (teaching, modeling, think-alouds, conferences, and/or organizers) is provided for aspects of the task that are being scored with the rubric

Student Instructions:

- Fully describes all student expectations.
- Partially describes student expectations.
- Minimally describes student expectations.

Comments:



Content Area: _____ Name of Assessment: _____
 Grade Level: _____ Date of Review: _____
 Reviewer (s): _____

PACE High Quality Assessment Review Tool

A high quality teacher-created assessment should be ... Aligned	
Part 2: Alignment	
	<p>The standards evaluated by the assessment are identified and are aligned to the expectations of the task:</p> <p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> Partial/Unclear</p> <p><input type="checkbox"/> No</p>
	<p>The standards and objectives are appropriate for the intended grade level that the assessment is being used for?</p> <p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> Partial/Unclear</p> <p><input type="checkbox"/> No</p>
	<p>The skills and knowledge assessed are grade level appropriate:</p> <p><input type="checkbox"/> Yes</p> <p><input type="checkbox"/> Partial/Unclear</p> <p><input type="checkbox"/> No</p>
	<p>To what extent do you see a content match between the prompt on the task and the corresponding Standards?</p> <p><input type="checkbox"/> Full match – all aspects of the task or items fully address or exceed the relevant skills and knowledge described in the corresponding standard(s)</p> <p><input type="checkbox"/> Close match – most aspects of the task or items address the relevant skills and knowledge described in the corresponding state standard(s)</p> <p><input type="checkbox"/> Partial match – Some aspects of the task or items address or partially address the skills and knowledge described in the corresponding state standard(s)</p> <p><input type="checkbox"/> Minimal match – Few aspects of the task or items match some relevant skills and knowledge described in the corresponding state standard(s)</p> <p><input type="checkbox"/> No match – No aspects of the task or items are related to the skills and knowledge described in the corresponding state standard(s)</p>



Content Area: _____ **Name of Assessment:** _____
Grade Level: _____ **Date of Review:** _____
Reviewer (s): _____

PACE High Quality Assessment Review Tool

Identify the Depth-of-Knowledge range of the Standards measured by the assessment (see Webb's DOK charts): <input type="checkbox"/> DOK 1: recall and reproduction <input type="checkbox"/> DOK 2: skills and concepts <input type="checkbox"/> DOK 3: strategic thinking/reasoning; requires deeper cognitive processing. <input type="checkbox"/> DOK 4: extended thinking; requires higher-order thinking including complex reasoning, planning, and developing of concepts.
Are the set of items or tasks reviewed as cognitively challenging as the standards? In other words, the student performance task elicits sufficient evidence for judging the level of student understanding related to the competencies and standards identified. Use the definitions below to select your rating: <input type="checkbox"/> More rigor – most items or the tasks reviewed are at a higher DOK level than the range indicated for the state standard(s) <input type="checkbox"/> Similar rigor – most items or the task reviewed are similar to the DOK range indicated for the state standard(s) <input type="checkbox"/> Less rigor – most items or the task reviewed are lower than the DOK range indicated for the state standard(s)
Comments/Suggestions for Improving Alignment (if any)
Relevant evidence to justify ratings:



Content Area: _____ Name of Assessment: _____
 Grade Level: _____ Date of Review: _____
 Reviewer (s): _____

PACE High Quality Assessment Review Tool

A high quality assessment should be ... Scored using Clear Guidelines and Criteria	
Part 3: Rubric	
PACE Rubric is used for the assessment:	<input type="checkbox"/> Yes <input type="checkbox"/> Earlier version <input type="checkbox"/> No
Other Content Rubric used for the assessment:	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is the rubric are aligned to the assessment task?	<input type="checkbox"/> Fully aligned <input type="checkbox"/> Partially aligned <input type="checkbox"/> Not aligned
Are the score categories clearly defined and coherent across performance levels?	<input type="checkbox"/> Yes <input type="checkbox"/> Partial <input type="checkbox"/> No
Is it clear which aspects of the task this rubric will be used to evaluate?	<input type="checkbox"/> Yes <input type="checkbox"/> Partial/Unclear <input type="checkbox"/> No
Based on your review of the rubric would the scoring rubric would most likely lead different raters to arrive at the same score for a given response?	<input type="checkbox"/> Yes <input type="checkbox"/> Partial/Unclear <input type="checkbox"/> No



Content Area: _____ Name of Assessment: _____
 Grade Level: _____ Date of Review: _____
 Reviewer (s): _____

PACE High Quality Assessment Review Tool

A high quality performance assessment should be...Fair and Unbiased	
Part 4: Fair and Unbiased	
(the areas below should be discussed relative to the needs of ELLs, gifted and talented students, and students with disabilities)	
To what extent are the tasks visually clear and uncluttered (e.g., appropriate white space and/or lines for student responses, graphics and/or illustrations are clear and support the test content, the font size seems appropriate for the students)?	
<input type="checkbox"/> Formatting is visually clear and uncluttered <input type="checkbox"/> Formatting is somewhat confusing or distracting <input type="checkbox"/> Formatting is unclear, cluttered, and inappropriate for students	
Are the directions and the task presented in as straightforward a way as possible for a range of learners?	
<input type="checkbox"/> Yes <input type="checkbox"/> Partial/Unclear <input type="checkbox"/> No	
Is the vocabulary and context(s) presented by the task free from cultural or other unintended bias?	
<input type="checkbox"/> Yes <input type="checkbox"/> Partial/Unclear <input type="checkbox"/> No	
Comments/Suggestions for Improvement for Fair and Unbiased (if any)	
Relevant evidence to justify ratings:	



Content Area: _____ Name of Assessment: _____
 Grade Level: _____ Date of Review: _____
 Reviewer (s): _____

PACE High Quality Assessment Review Tool

A high quality performance assessment includes appropriate reading and visual materials	
Part 5: Appropriateness of Text/Visual Resources	
The texts and visual resources support the topic and prompt:	
<input type="checkbox"/> Yes <input type="checkbox"/> Partial/Unclear <input type="checkbox"/> No <input type="checkbox"/> N/A	
The texts have characteristics of a:	
<input type="checkbox"/> Simple Text <input type="checkbox"/> Somewhat Complex Texts <input type="checkbox"/> Complex Texts <input type="checkbox"/> Very Complex Texts <input type="checkbox"/> N/A Note: Refer to the <i>Text Complexity Rubric for Literary Texts or Informational Texts</i>	
The amount of texts and visual resources are:	
<input type="checkbox"/> Appropriate for the grade level and the time allotted for the task <input type="checkbox"/> Appropriate for the grade level, but may exceed the time allotted for the task <input type="checkbox"/> Burdensome for the grade level and the time allotted for the task <input type="checkbox"/> No texts and/or resources are included <input type="checkbox"/> N/A	
Comments/Suggestions for Improvement for Fair and Unbiased (if any)	
Relevant evidence to justify ratings:	



Content Area: _____ **Name of Assessment:** _____
Grade Level: _____ **Date of Review:** _____
Reviewer (s): _____

PACE High Quality Assessment Review Tool

Recommendation for this assessment:

- No changes needed**
- Minor changes recommended**
- Some changes required, please address and resubmit**
- Substantial changes needed, please address and resubmit**
- Task rejected—new task needed**

Discussion:



APPENDIX C: CRITERIA FOR PROCURING AND EVALUATING HIGH-QUALITY ASSESSMENTS (CCSSO, 2014)

A. Meet Overall Assessment Goals and Ensure Technical Quality

- A.1 Indicating progress toward college and career readiness
- A.2 Ensuring that assessments are valid for required and intended purposes
- A.3 Ensuring that assessments are reliable
- A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years
- A.5 Providing accessibility to all students, including English learners and students with disabilities
- A.6 Ensuring transparency of test design and expectations
- A.7 Meeting all requirements for data privacy and ownership

B. Align to Standards – English Language Arts/Literacy

- B.1 Assessing student reading and writing achievement in both ELA and literacy
- B.2 Focusing on complexity of texts
- B.3 Requiring students to read closely and use evidence from texts
- B.4 Requiring a range of cognitive demand
- B.5 Assessing writing
- B.6 Emphasizing vocabulary and language skills
- B.7 Assessing research and inquiry
- B.8 Assessing speaking and listening
- B.9 Ensuring high-quality items and a variety of item types

C. Align to Standards – Mathematics

- C.1 Focusing strongly on the content most needed for success in later mathematics
- C.2 Assessing a balance of concepts, procedures, and applications
- C.3 Connecting practice to content
- C.4 Requiring a range of cognitive demand
- C.5 Ensuring high-quality items and a variety of item types



D. Yield Valuable Reports on Student Progress and Performance

- D.1 Focusing on student achievement and progress to readiness
- D.2 Providing timely data that inform instruction

E. Adhere to Best Practices in Test Administration

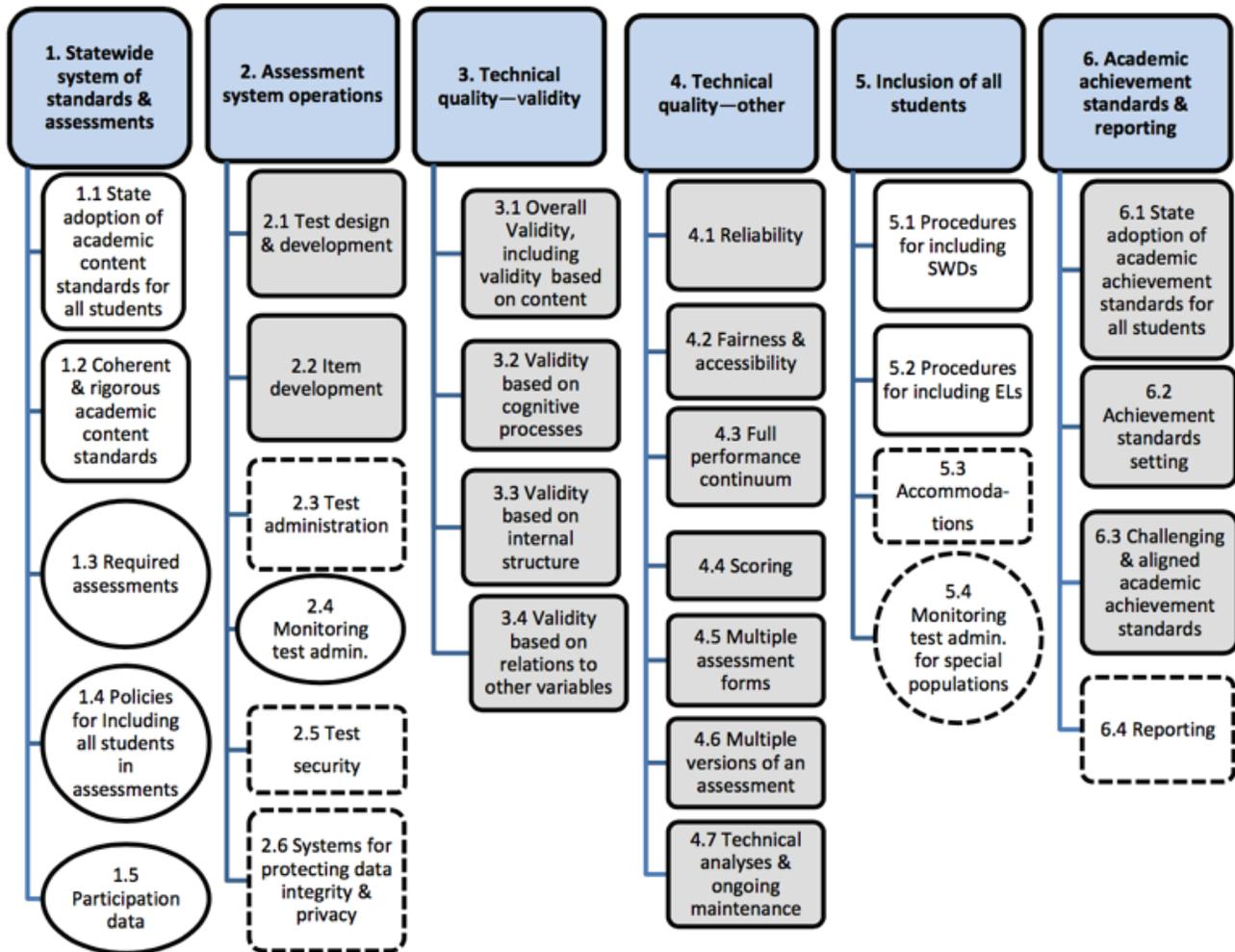
- E.1 Maintaining necessary standardization and ensuring test security

F. State Specific Criteria (as desired)

Sample criteria might include

- Requiring involvement of the state’s K-12 educators and institutions of higher education
- Procuring a system of aligned assessments, including diagnostic and interim assessments
- Ensuring interoperability of computer-administered items

APPENDIX D: MAP OF THE CRITICAL ELEMENTS FOR THE STATE ASSESSMENT PEER REVIEW (USED, 2015)



KEY

- Critical elements in ovals will be checked for completeness by Department staff; if necessary, they may also be reviewed by assessment peer reviewers (e.g., Critical Element 1.3). All other critical elements will be reviewed by assessment peer reviewers.
- Critical elements in shaded boxes likely will be addressed by coordinated evidence for all States administering the same assessments (e.g., Critical Element 2.1).
- Critical elements in clear boxes with solid outlines likely will be addressed with State-specific evidence, even if a State administers the same assessments administered by other States (e.g., Critical Element 5.1).
- Critical elements in ovals or clear boxes with dashed outlines likely will be addressed by both State-specific evidence and coordinated evidence for States administering the same assessments (e.g., Critical Element 2.3, 5.4).

APPENDIX E: ASSESSMENT PEER REVIEW GUIDANCE EXAMPLES OF EVIDENCE FOR CRITICAL ELEMENTS

Assessment Peer Review Guidance

U.S. Department of Education

SECTION 1: STATEWIDE SYSTEM OF STANDARDS AND ASSESSMENTS

Critical Element 1.1 – State Adoption of Academic Content Standards for All Students

	Examples of Evidence
<p>The State formally adopted challenging academic content standards for all students in reading/language arts, mathematics and science and applies its academic content standards to all public elementary and secondary schools and students in the State.</p>	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • Evidence of adoption of the State’s academic content standards, specifically: <ul style="list-style-type: none"> ○ Indication of <i>Requirement Previously Met</i>; or ○ State Board of Education minutes, memo announcing formal approval from the Chief State School Officer to districts, legislation, regulations, or other binding approval of a particular set of academic content standards; • Documentation, such as text prefacing the State’s academic content standards, policy memos, State newsletters to districts, or other key documents, that explicitly state that the State’s academic content standards apply to all public elementary and secondary schools and all public elementary and secondary school students in the State; <p>Note: A State with <i>Requirement Previously Met</i> should note the applicable category in the State Assessment Peer Review Submission Index for its peer submission. <i>Requirement Previously Met</i> applies to a State in the following categories: (1) a State that has academic content standards in reading/language arts, mathematics, or science that have not changed significantly since the State’s previous assessment peer review; or (2) with respect to academic content standards in reading/language arts and mathematics, a State approved for ESEA flexibility that (a) has adopted a set of college- and career-ready academic content standards that are common to a significant number of States and has not adopted supplemental State-specific academic content standards in these content areas, or (b) has adopted a set of college- and career-ready academic content standards certified by a State network of institutions of higher education (IHEs).</p>

Critical Element 1.2 – Coherent and Rigorous Academic Content Standards

	Examples of Evidence
<p>The State’s academic content standards in reading/language arts, mathematics and science specify what students are expected to know and be able to do by the time they graduate from high school to succeed in college and the workforce; contain content that is coherent (e.g., within and across grades) and rigorous; encourage the teaching of advanced skills; and were developed with broad stakeholder involvement.</p>	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • Indication of <i>Requirement Previously Met</i>; or • Evidence that the State’s academic content standards: <ul style="list-style-type: none"> ○ Contain coherent and rigorous content and encourage the teaching of advanced skills, such as: <ul style="list-style-type: none"> ▪ A detailed description of the strategies the State used to ensure that its academic content standards adequately specify what students should know and be able to do; ▪ Documentation of the process used by the State to benchmark its academic content standards to nationally or internationally recognized academic content standards; ▪ Reports of external independent reviews of the State’s academic content standards by content experts, summaries of reviews by educators in the State, or other documentation to confirm that the State’s academic content standards adequately specify what students should know and be able to do;



	<ul style="list-style-type: none"> ▪ Endorsements or certifications by the State’s network of institutions of higher education (IHEs), professional associations and/or the business community that the State’s academic content standards represent the knowledge and skills in the content area(s) under review necessary for students to succeed in college and the workforce; ○ Were developed with broad stakeholder involvement, such as: <ul style="list-style-type: none"> ▪ Summary report of substantive involvement and input of educators, such as committees of curriculum, instruction, and content specialists, teachers and others, in the development of the State’s academic content standards; ▪ Documentation of substantial involvement of subject-matter experts, including teachers, in the development of the State’s academic content standards; ▪ Descriptions that demonstrate a broad range of stakeholders was involved in the development of the State’s academic content standards, including individuals representing groups such as students with disabilities, English learners and other student populations in the State; parents; and the business community; ▪ Documentation of public hearings, public comment periods, public review, or other activities that show broad stakeholder involvement in the development or adoption of the State’s academic content standards. <p>Note: See note in Critical Element 1.1 – State Adoption of Academic Content Standards for All Students. With respect to academic content standards in reading/language arts and mathematics, <i>Requirement Previously Met</i> does not apply to supplemental State-specific academic content standards for a State approved for ESEA flexibility that has adopted a set of college- and career-ready academic content standards in a content area that are common to a significant number of States and also adopted supplemental State-specific academic content standards in that content area.</p>
--	--



Critical Element 1.3 – Required Assessments

	Examples of Evidence
<p>The State’s assessment system includes annual general and alternate assessments (based on grade-level academic achievement standards or alternate academic achievement standards) in:</p> <ul style="list-style-type: none"> • Reading/language arts and mathematics in each of grades 3-8 and at least once in high school (grades 10-12); • Science at least once in each of three grade spans (3-5, 6-9 and 10-12). 	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • A list of the annual assessments the State administers in reading/language arts, mathematics and science including, as applicable, alternate assessments based on grade-level academic achievement standards or alternate academic achievement standards for students with the most significant cognitive disabilities, and native language assessments, and the grades in which each type of assessment is administered.

Critical Element 1.4 – Policies for Including All Students in Assessments

	Examples of Evidence
<p>The State requires the inclusion of all public elementary and secondary school students in its assessment system and clearly and consistently communicates this requirement to districts and schools.</p> <ul style="list-style-type: none"> • For students with disabilities, policies state that all students with disabilities in the State, including students with disabilities publicly placed in private schools as a means of providing special education and related services, must be included in the assessment system; • For English learners: <ul style="list-style-type: none"> ○ Policies state that all English learners must be included in the assessment system, unless the State exempts a student who has attended schools in the U.S. for less than 12 months from one administration of its reading/ 	<p>Evidence to support this critical element for the State’s assessment system includes documents such as:</p> <ul style="list-style-type: none"> • Key documents, such as regulations, policies, procedures, test coordinator manuals, test administrator manuals and accommodations manuals that the State disseminates to educators (districts, schools and teachers), that clearly state that all students must be included in the State’s assessment system and do not exclude any student group or subset of a student group; • For students with disabilities, if needed to supplement the above: Instructions for Individualized Education Program (IEP) teams and/or other key documents; • For English learners, if applicable and needed to supplement the above: Test administrator manuals and/or other key documents that show that the State provides a native language (e.g., Spanish, Vietnamese) version of its assessments.



<p>language arts assessment;</p> <ul style="list-style-type: none"> ○ If the State administers native language assessments, the State requires English learners to be assessed in reading/language arts in English if they have been enrolled in U.S. schools for three or more consecutive years, except if a district determines, on a case-by-case basis, that native language assessments would yield more accurate and reliable information, the district may assess a student with native language assessments for a period not to exceed two additional consecutive years. 	
--	--

Critical Element 1.5 – Participation Data

	Examples of Evidence
<p>The State’s participation data show that all students, disaggregated by student group and assessment type, are included in the State’s assessment system. In addition, if the State administers end-of-course assessments for high school students, the State has procedures in place for ensuring that each student is tested and counted in the calculation of participation rates on each required assessment and provides the corresponding data.</p>	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • Participation data from the most recent year of test administration in the State, such as in Table 1 below, that show that all students, disaggregated by student group (i.e., students with disabilities, English learners, economically disadvantaged students, students in major racial/ethnic categories, migratory students, and male/female students) and assessment type (i.e., general and AA-AAAS) in the tested grades are included in the State’s assessments for reading/language arts, mathematics and science; • If the State administers end-of-course assessments for high school students, evidence that the State has procedures in place for ensuring that each student is included in the assessment system during high school, including students with the most significant cognitive disabilities who take an alternate assessment based on alternate academic achievement standards and recently arrived English learners who take an ELP assessment in lieu of a reading/language arts assessment, such as: <ul style="list-style-type: none"> ○ Description of the method used for ensuring that each student is tested and counted in the calculation of participation rate on each required assessment. If course enrollment or another proxy is used to count all students, a description of the method used to ensure that all students are counted in the proxy measure; ○ Data that reflect implementation of participation rate calculations that ensure that each student is tested and counted for each required assessment. Also, if course enrollment or another proxy is used to count all students, data that document that all students are counted in the proxy measure.



Table 1: Students Tested by Student Group in [subject] during [school year]								
Student group		Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	HS
All	# enrolled							
	# tested							
	% tested							
Economically disadvantaged	# enrolled							
	# tested							
	% tested							
Students with disabilities	# enrolled							
	# tested							
	% tested							
(Continued for all other student groups)	# enrolled							
	# tested							
	% tested							

Number of students assessed on the State’s AA-AAAS in [subject] during [school year]: _____.

Note: A student with a disability should only be counted as tested if the student received a valid score on the State’s general or alternate assessments submitted for assessment peer review for the grade in which the student was enrolled. If the State permits a recently arrived English learner (i.e., an English learner who has attended schools in the U.S. for less than 12 months) to be exempt from one administration of the State’s reading/language arts assessment and to take the State’s ELP assessment in lieu of the State’s reading/language arts assessment, then the State should count such students as ‘tested’ in data submitted to address critical element.



SECTION 2: ASSESSMENT SYSTEM OPERATIONS

Critical Element 2.1 – Test Design and Development

	Examples of Evidence
<p>The State’s test design and test development process is well-suited for the content, is technically sound, aligns the assessments to the full range of the State’s academic content standards, and includes:</p> <ul style="list-style-type: none"> • Statement(s) of the purposes of the assessments and the intended interpretations and uses of results; • Test blueprints that describe the structure of each assessment in sufficient detail to support the development of assessments that are technically sound, measure the full range of the State’s grade-level academic content standards, and support the intended interpretations and uses of the results; • Processes to ensure that each assessment is tailored to the knowledge and skills included in the State’s academic content standards, reflects appropriate inclusion of challenging content, and requires complex demonstrations or applications of knowledge and skills (i.e., higher-order thinking skills); • If the State administers computer-adaptive assessments, the item pool and item selection procedures adequately support the test design. 	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes:</p> <p>For the State’s general assessments:</p> <ul style="list-style-type: none"> • Relevant sections of State code or regulations, language from contract(s) for the State’s assessments, test coordinator or test administrator manuals, or other relevant documentation that states the purposes of the assessments and the intended interpretations and uses of results; • Test blueprints that: <ul style="list-style-type: none"> ○ Describe the structure of each assessment in sufficient detail to support the development of a technically sound assessment, for example, in terms of the number of items, item types, the proportion of item types, response formats, range of item difficulties, types of scoring procedures, and applicable time limits; ○ Align to the State’s grade-level academic content standards in terms of content (i.e. knowledge and cognitive process), the full range of the State’s grade-level academic content standards, balance of content, and cognitive complexity; • Documentation that the test design that is tailored to the specific knowledge and skills in the State’s academic content standards (e.g., includes extended response items that require demonstration of writing skills if the State’s reading/language arts academic content standards include writing); • Documentation of the approaches the State uses to include challenging content and complex demonstrations or applications of knowledge and skills (i.e., items that assess higher-order thinking skills, such as item types appropriate to the content that require synthesizing and evaluating information and analytical text-based writing or multiple steps and student explanations of their work); for example, this could include test specifications or test blueprints that require a certain portion of the total score be based on item types that require complex demonstrations or applications of knowledge and skills and the rationale for that design. <p> For the State’s technology-based general assessments, in addition to the above:</p> <ul style="list-style-type: none"> • Evidence of the usability of the technology-based presentation of the assessments, including the usability of accessibility tools and features (e.g., embedded in test items or available as an accompaniment to the items), such as descriptions of conformance with established accessibility standards and best practices and usability studies; • For computer-adaptive general assessments: <ul style="list-style-type: none"> ○ Evidence regarding the item pool, including: <ul style="list-style-type: none"> ▪ Evidence regarding the size of the item pool and the characteristics (non-statistical (e.g., content) and statistical) of the items it contains that demonstrates that the item pool has the capacity to produce test forms that adequately reflect the State’s test blueprints in terms of: <ul style="list-style-type: none"> - Full range of the State’s academic content standards, balance of content, cognitive complexity for



	<p>each academic content standard, and range of item difficulty levels for each academic content standard;</p> <ul style="list-style-type: none"> - Structure of the assessment (e.g., numbers of items, proportion of item types and response types); <ul style="list-style-type: none"> o Technical documentation for item selection procedures that includes descriptive evidence and empirical evidence (e.g., simulation results that reflect variables such as a wide range of student behaviors and abilities and test administration early and late in the testing window) that show that the item selection procedures are designed adequately for: <ul style="list-style-type: none"> ▪ Content considerations to ensure test forms that adequately reflect the State’s academic content standards in terms of the full range of the State’s grade-level academic content standards, balance of content, and the cognitive complexity for each standard tested; ▪ Structure of the assessment specified by the blueprints; ▪ Reliability considerations such that the test forms produce adequately precise estimates of student achievement for all students (e.g., for students with consistent and inconsistent testing behaviors, high- and low-achieving students; English learners and students with disabilities); ▪ Routing students appropriately to the next item or stage; ▪ Other operational considerations, including starting rules (i.e., selection of first item), stopping rules, and rules to limit item over-exposure. <p>AA-AAAS. For the State’s AA-AAAS:</p> <ul style="list-style-type: none"> • Relevant sections of State code or regulations, language from contract(s) for the State’s assessments, test coordinator or test administrator manuals, or other relevant documentation that states the purposes of the assessments and the intended interpretations and uses of results for students tested; • Description of the structure of the assessment, for example, in terms of the number of items, item types, the proportion of item types, response formats, types of scoring procedures, and applicable time limits. For a portfolio assessment, the description should include the purpose and design of the portfolio, exemplars, artifacts, and scoring rubrics; • Test blueprints (or, where applicable, specifications for the design of portfolio assessments) that reflect content linked to the State’s grade-level academic content standards and the intended breadth and cognitive complexity of the assessments; • To the extent the assessments are designed to cover a narrower range of content than the State’s general assessments and differ in cognitive complexity: <ul style="list-style-type: none"> o Description of the breadth of the grade-level academic content standards the assessments are designed to measure, such as an evidence-based rationale for the reduced breadth within each grade and/or comparison of intended content compared to grade-level academic content standards; o Description of the strategies the State used to ensure that the cognitive complexity of the assessments is appropriately challenging for students with the most significant cognitive disabilities; o Description of how linkage to different content across grades/grade spans and vertical articulation of academic expectations for students is maintained; • If the State developed extended academic content standards to show the relationship between the State’s grade-level academic content standards and the content of the assessments, documentation of their use in the design
--	---



	<p>of the assessments;</p> <ul style="list-style-type: none"> • For adaptive alternate assessments (both computer-delivered and human-delivered), evidence, such as a technical report for the assessments, showing: <ul style="list-style-type: none"> ○ Evidence that the size of the item pool and the characteristics of the items it contains are appropriate for the test design; ○ Evidence that rules in place for routing students are designed to produce test forms that adequately reflect the blueprints and produce adequately precise estimates of student achievement for classifying students; ○ Evidence that the rules for routing students, including starting (e.g., selection of first item) and stopping rules, are appropriate and based on adequately precise estimates of student responses, and are not primarily based on the effects of a student’s disability, including idiosyncratic knowledge patterns; • For technology-based AA-AAAS, in addition to the above, evidence of the usability of the technology-based presentation of the assessments, including the usability of accessibility tools and features (e.g., embedded in test items or available as an accompaniment to the items), such as descriptions of conformance with established accessibility standards and best practices and usability studies.
--	--

Critical Element 2.2 – Item Development

	Examples of Evidence
<p>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.</p>	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes documents such as:</p> <p>For the State’s general assessments, evidence, such as a sections in the technical report for the assessments, that show:</p> <ul style="list-style-type: none"> • A description of the process the State uses to ensure that the item types (e.g., multiple choice, constructed response, performance tasks, and technology-enhanced items) are tailored for assessing the academic content standards in terms of content; • A description of the process the State uses to ensure that items are tailored for assessing the academic content standards in terms of cognitive process (e.g., assessing complex demonstrations of knowledge and skills appropriate to the content, such as with item types that require synthesizing and evaluating information and analytical text-based writing or multiple steps and student explanations of their work); • Samples of item specifications that detail the content standards to be tested, item type, intended cognitive complexity, intended level of difficulty, accessibility tools and features, and response format; • Description or examples of instructions provided to item writers and reviewers; • Documentation that items are developed by individuals with content area expertise, experience as educators, and experience and expertise with students with disabilities, English learners, and other student populations in the State; • Documentation of procedures to review items for alignment to academic content standards, intended levels of cognitive complexity, intended levels of difficulty, construct-irrelevant variance, and consistency with item specifications, such as documentation of content and bias reviews by an external review committee;



	<ul style="list-style-type: none"> • Description of procedures to evaluate the quality of items and select items for operational use, including evidence of reviews of pilot and field test data; • As applicable, evidence that accessibility tools and features (e.g., embedded in test items or available as an accompaniment to the items) do not produce an inadvertent effect on the construct assessed; • Evidence that the items elicit the intended response processes, such as cognitive labs or interaction studies. <p>AA-AAAS. For the State’s AA-AAAS, in addition to the above:</p> <ul style="list-style-type: none"> • If the State’s AA-AAAS is a portfolio assessment, samples of item specifications that include documentation of the requirements for student work and samples of exemplars for illustrating levels of student performance; • Documentation of the process the State uses to ensure that the assessment items are accessible, cognitively challenging, and reflect professional judgment of the highest achievement standards possible for students with the most significant cognitive disabilities. <p> For the State’s technology-based general assessments and AA-AAAS:</p> <ul style="list-style-type: none"> • Documentation that procedures to evaluate and select items considered the deliverability of the items (e.g., usability studies). <p>Note: This critical element is closely related to Critical Element 4.2 – Fairness and Accessibility.</p>
--	---

Critical Element 2.3 – Test Administration

	Examples of Evidence
<p>The State implements policies and procedures for standardized test administration, specifically the State:</p> <ul style="list-style-type: none"> • Has established and communicates to educators clear, thorough and consistent standardized procedures for the administration of its assessments, including administration with accommodations; • Has established procedures to ensure that all individuals responsible for administering the State’s general and alternate assessments receive training on the State’s established procedures for the administration of its assessments; 	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes:</p> <ul style="list-style-type: none"> • Regarding test administration: <ul style="list-style-type: none"> ○ Test coordinator manuals, test administration manuals, accommodations manuals and/or other key documents that the State provides to districts, schools, and teachers that address standardized test administration and any accessibility tools and features available for the assessments; ○ Instructions for the use of accommodations allowed by the State that address each accommodation. For example: <ul style="list-style-type: none"> ▪ For accommodations such as bilingual dictionaries for English learners, instructions that indicate which types of bilingual dictionaries are and are not acceptable and how to acquire them for student use during the assessment; ▪ For accommodations such as readers and scribes for students with disabilities, documentation of expectations for training and test security regarding test administration with readers and scribes; ○ Evidence that the State provides key documents regarding test administration to district and school test coordinators and administrators, such as e-mails, websites, or listserv messages to inform relevant staff of the availability of documents for downloading or cover memos that accompany hard copies of the materials



<ul style="list-style-type: none"> • If the State administers technology-based assessments, the State has defined technology and other related requirements, included technology-based test administration in its standardized procedures for test administration, and established contingency plans to address possible technology challenges during test administration. 	<p>delivered to districts and schools;</p> <ul style="list-style-type: none"> ○ Evidence of the State’s process for documenting modifications or disruptions of standardized test administration procedures (e.g., unapproved non-standard accommodations, electric power failures or hardware failures during technology-based testing), such as sample of incidences documented during the most recent year of test administration in the State. <ul style="list-style-type: none"> • Regarding training for test administration: <ul style="list-style-type: none"> ○ Evidence regarding training, such as: <ul style="list-style-type: none"> ▪ Schedules for training sessions for different groups of individuals involved in test administration (e.g., district and school test coordinators, test administrators, school computer lab staff, accommodation providers); ▪ Training materials, such as agendas, slide presentations and school test coordinator manuals and test administrator manuals, provided to participants. For technology-based assessments, training materials that include resources such as practice tests and/or other supports to ensure that test coordinators, test administrators and others involved in test administration are prepared to administer the assessments; ▪ Documentation of the State’s procedures to ensure that all test coordinators, test administrators, and other individuals involved in test administration receive training for each test administration, such as forms for sign-in sheets or screenshots of electronic forms for tracking attendance, assurance forms, or identification of individuals responsible for tracking attendance. <p> For the State’s technology-based assessments:</p> <ul style="list-style-type: none"> • Evidence that the State has clearly defined the technology (e.g., hardware, software, internet connectivity, and internet access) and other related requirements (e.g., computer lab configurations) necessary for schools to administer the assessments and has communicated these requirements to schools and districts; • District and school test coordinator manuals, test administrator manuals and/or other key documents that include specific instructions for administering technology-based assessments (e.g., regarding necessary advanced preparation, ensuring that test administrators and students are adequately familiar with the delivery devices and, as applicable, accessibility tools and features available for students); • Contingency plans or summaries of contingency plans that outline strategies for managing possible challenges or disruptions during test administration. <p>AA-AAAS. For the State’s AA-AAAS, in addition to the above:</p> <ul style="list-style-type: none"> • If the assessments involve teacher-administered performance tasks or portfolios, key documents, such as test administration manuals, that the State provides to districts, schools and teachers that include clear, precise descriptions of activities, standard prompts, exemplars and scoring rubrics, as applicable; and standard procedures for the administration of the assessments that address features such as determining entry points, selection and use of manipulatives, prompts, scaffolding, and recognizing and recording responses; • Evidence that training for test administrators addresses key assessment features, such as teacher-administered performance tasks or portfolios; determining entry points; selection and use of manipulatives; prompts;
---	---



	scaffolding; recognizing and recording responses; and/or other features for which specific instructions may be needed to ensure standardized administration of the assessment.
--	--

Critical Element 2.4 – Monitoring Test Administration

	Examples of Evidence
The State adequately monitors the administration of its State assessments to ensure that standardized test administration procedures are implemented with fidelity across districts and schools.	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes documents such as:</p> <ul style="list-style-type: none"> • Brief description of the State’s approach to monitoring test administration (e.g., monitoring conducted by State staff, through regional centers, by districts with support from the State, or another approach); • Existing written documentation of the State’s procedures for monitoring test administration across the State, including, for example, strategies for selection of districts and schools for monitoring, cycle for reaching schools and districts across the State, schedule for monitoring, monitors’ roles, and the responsibilities of key personnel; • Summary of the results of the State’s monitoring of the most recent year of test administration in the State.

Critical Element 2.5 – Test Security

	Examples of Evidence
<p>The State has implemented and documented an appropriate set of policies and procedures to prevent test irregularities and ensure the integrity of test results through:</p> <ul style="list-style-type: none"> • Prevention of any assessment irregularities, including maintaining the security of test materials, proper test preparation guidelines and administration procedures, incident-reporting procedures, consequences for confirmed violations of test security, and requirements for annual training at the district and school levels for all individuals involved in test administration; • Detection of test irregularities; • Remediation following any test security incidents involving any of 	<p>Collectively, evidence to support this critical element for the State’s assessment system must demonstrate that the State has implemented and documented an appropriate approach to test security.</p> <p>Evidence to support this critical element for the State’s assessment system may include:</p> <ul style="list-style-type: none"> • State Test Security Handbook; • Summary results or reports of internal or independent monitoring, audit, or evaluation of the State’s test security policies, procedures and practices, if any. <p>Evidence of procedures for prevention of test irregularities includes documents such as:</p> <ul style="list-style-type: none"> • Key documents, such as test coordinator manuals or test administration manuals for district and school staff, that include detailed security procedures for before, during and after test administration; • Documented procedures for tracking the chain of custody of secure materials and for maintaining the security of test materials at all stages, including distribution, storage, administration, and transfer of data; • Documented procedures for mitigating the likelihood of unauthorized communication, assistance, or recording of test materials (e.g., via technology such as smart phones); • Specific test security instructions for accommodations providers (e.g., readers, sign language interpreters, special education teachers and support staff if the assessment is administered individually), as applicable; • Documentation of established consequences for confirmed violations of test security, such as State law, State



<p>the State’s assessments;</p> <ul style="list-style-type: none"> Investigation of alleged or factual test irregularities. 	<p>regulations or State Board-approved policies;</p> <ul style="list-style-type: none"> Key documents such as policy memos, listserv messages, test coordinator manuals and test administration manuals that document that the State communicates its test security policies, including consequences for violation, to all individuals involved in test administration; Newsletters, listserv messages, test coordinator manuals, test administrator manuals and/or other key documents from the State that clearly state that annual test security training is required at the district and school levels for all staff involved in test administration; Evidence submitted under Critical Element 2.3 – Test Administration that shows: <ul style="list-style-type: none"> The State’s test administration training covers the relevant aspects of the State’s test security policies; Procedures for ensuring that all individuals involved in test administration receive annual test security training. <p> For the State’s technology-based assessments, evidence of procedures for prevention of test irregularities includes:</p> <ul style="list-style-type: none"> Documented policies and procedures for districts and schools to address secure test administration challenges related to hardware, software, internet connectivity, and internet access. <p>Evidence of procedures for detection of test irregularities includes documents such as:</p> <ul style="list-style-type: none"> Documented incident-reporting procedures, such as a template and instructions for reporting test administration irregularities and security incidents for district, school and other personnel involved in test administration; Documentation of the information the State routinely collects and analyzes for test security purposes, such as description of post-administration data forensics analysis the State conducts (e.g., unusual score gains or losses, similarity analyses, erasure/answer change analyses, pattern analysis, person fit analyses, local outlier detection, unusual timing patterns); Summary of test security incidents from most recent year of test administration (e.g., types of incidents and frequency) and examples of how they were addressed, or other documentation that demonstrates that the State identifies, tracks, and resolves test irregularities. <p>Evidence of procedures for remediation of test irregularities includes documents such as:</p> <ul style="list-style-type: none"> Contingency plan that demonstrates that the State has a plan for how to respond to test security incidents and that addresses: <ul style="list-style-type: none"> Different types of possible test security incidents (e.g., human, physical, electronic, or internet-related), including those that require immediate action (e.g., items exposed on-line during the testing window); Policies and procedures the State would use to address different types of test security incidents (e.g., continue vs. stop testing, retesting, replacing existing forms or items, excluding items from scoring, invalidating results); Communication strategies for communicating with districts, schools and others, as appropriate, for addressing active events.
--	--



	<p>Evidence of procedures for investigation of alleged or factual test irregularities includes documents such as:</p> <ul style="list-style-type: none"> • State’s policies and procedures for responding to and investigating, where appropriate, alleged or actual security lapses and test irregularities that: <ul style="list-style-type: none"> ○ Include securing evidence in cases where an investigation may be pursued; ○ Include the State’s decision rules for investigating potential test irregularities; ○ Provide standard procedures and strategies for conducting investigations, including guidelines to districts, if applicable; ○ Include policies and procedures to protect the privacy and professional reputation of all parties involved in an investigation. <p>Note: Evidence should be redacted to protect personally identifiable information, as appropriate.</p>
--	--

Critical Element 2.6 – Systems for Protecting Data Integrity and Privacy

	Examples of Evidence
<p>The State has policies and procedures in place to protect the integrity and confidentiality of its test materials, test-related data, and personally identifiable information, specifically:</p> <ul style="list-style-type: none"> • To protect the integrity of its test materials and related data in test development, administration, and storage and use of results; • To secure student-level assessment data and protect student privacy and confidentiality, including guidelines for districts and schools; • To protect personally identifiable information about any individual student in reporting, including defining the minimum number of students necessary to allow reporting of scores for all students and student groups. 	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes documents such as:</p> <ul style="list-style-type: none"> • Evidence of policies and procedures to protect the integrity and confidentiality of test materials and test-related data, such as: <ul style="list-style-type: none"> ○ State security plan, or excerpts from the State’s assessment contracts or other materials that show expectations, rules and procedures for reducing security threats and risks and protecting test materials and related data during item development, test construction, materials production, distribution, test administration, and scoring; ○ Description of security features for storage of test materials and related data (i.e., items, tests, student responses, and results); ○ Rules and procedures for secure transfer of student-level assessment data in and out of the State’s data management and reporting systems; between authorized users (e.g., State, district and school personnel, and vendors); and at the local level (e.g., requirements for use of secure sites for accessing data, directions regarding the transfer of student data); ○ Policies and procedures for allowing only secure, authorized access to the State’s student-level data files for the State, districts, schools, and others, as applicable (e.g., assessment consortia, vendors); ○ Training requirements and materials for State staff, contractors and vendors, and others related to data integrity and appropriate handling of personally identifiable information; ○ Policies and procedures to ensure that aggregate or de-identified data intended for public release do not inadvertently disclose any personally identifiable information; ○ Documentation that the above policies and procedures, as applicable, are clearly communicated to all relevant personnel (e.g., State staff, assessment, districts, and schools, and others, as applicable (e.g., assessment consortia, vendors)); ○ Rules and procedures for ensuring that data released by third parties (e.g., agency partners, vendors,



	<p>external researchers) are reviewed for adherence to State Statistical Disclosure Limitation (SDL) standards and do not reveal personally identifiable information.</p> <ul style="list-style-type: none"> • Evidence of policies and procedures to protect personally identifiable information about any individual student in reporting, such as: <ul style="list-style-type: none"> ○ State operations manual or other documentation that clearly states the State’s SDL rules for determining whether data are reported for a group of students or a student group, including: <ul style="list-style-type: none"> ▪ Defining the minimum number of students necessary to allow reporting of scores for a student group; ▪ Rules for applying complementary suppression (or other SDL methods) when one or more student groups are not reported because they fall below the minimum reporting size; ▪ Rules for not reporting results, regardless of the size of the student group, when reporting would reveal personally identifiable information (e.g., procedures for reporting “<10%” for proficient and above when no student scored at those levels); ▪ Other rules to ensure that aggregate or de-identified data do not inadvertently disclose any personally identifiable information; ○ State operations manual or other document that describes how the State’s rules for protecting personally identifiable information are implemented.
--	--



SECTION 3: TECHNICAL QUALITY – VALIDITY

Critical Element 3.1 – Overall Validity, Including Validity Based on Content

	Examples of Evidence
<p>The State has documented adequate overall validity evidence for its assessments, and the State’s validity evidence includes evidence that the State’s assessments measure the knowledge and skills specified in the State’s academic content standards, including:</p> <ul style="list-style-type: none"> • Documentation of adequate alignment between the State’s assessments and the academic content standards the assessments are designed to measure in terms of content (i.e., knowledge and process), the full range of the State’s academic content standards, balance of content, and cognitive complexity; • If the State administers alternate assessments based on alternate academic achievement standards, the assessments show adequate linkage to the State’s academic content standards in terms of content match (i.e., no unrelated content) and the breadth of content and cognitive complexity determined in test design to be appropriate for students with the most significant cognitive disabilities. 	<p>Collectively, across the State’s assessments, evidence to support critical elements 3.1 through 3.4 for the State’s general assessments and AA-AAAS must document overall validity evidence generally consistent with expectations of current professional standards.</p> <p>Evidence to document adequate overall validity evidence for the State’s general assessments and AA-AAAS includes documents such as:</p> <ul style="list-style-type: none"> • A chapter on validity in the technical report for the State’s assessments that states the purposes of the assessments and intended interpretations and uses of results and shows validity evidence for the assessments that is generally consistent with expectations of current professional standards; • Other validity evidence, in addition to that outlined in critical elements 3.1 through 3.4, that is necessary to document adequate validity evidence for the assessments. <p>Evidence to document adequate validity evidence based on content for the State’s general assessments includes:</p> <ul style="list-style-type: none"> • Validity evidence based on the assessment content that shows levels of validity generally consistent with expectations of current professional standards, such as: <ul style="list-style-type: none"> ○ Test blueprints, as submitted under Critical Element 2.1—Test Design and Development; ○ A full form of the assessment in one grade for the general assessment in reading/language arts and mathematics (e.g., one form of the grade 5 mathematics assessment and one form of the grade 8 reading/language arts assessment),⁶ ○ Logical or empirical analyses that show that the test content adequately represents the full range of the State’s academic content standards; ○ Report of expert judgment of the relationship between components of the assessment and the State’s academic content standards; ○ Reports of analyses to demonstrate that the State’s assessment content is appropriately related to the specific inferences made from test scores about student proficiency in the State’s academic content standards for all student groups; • Evidence of alignment, including: <ul style="list-style-type: none"> ○ Report of results of an independent alignment study that is technically sound (i.e., method and process, appropriate units of analysis, clear criteria) and documents adequate alignment, specifically that: <ul style="list-style-type: none"> ▪ Each assessment is aligned to its test blueprint, and each blueprint is aligned to the full range of State’s

⁶ The Department recognizes the need for a State to maintain the security of its test forms; a State that elects to submit a test form(s) as part of its assessment peer review submission should contact the Department so that arrangements can be made to ensure that the security of the materials is maintained. Such materials will be reviewed by the assessment peer reviewers in accordance with the State’s test security requirements and agreements.



	<p>academic content standards; or</p> <ul style="list-style-type: none"> ▪ Each assessment is aligned to the full range of the State’s academic content standards, and the procedures the State follows to ensure such alignment during test development; ○ Description of a systematic process and timeline the State will implement to address any gaps or weaknesses identified in the alignment studies. <p> For the State’s computer-adaptive general assessments:</p> <ul style="list-style-type: none"> • Empirical evidence that the size of the item pool and the characteristics (non-statistical (e.g., content) and statistical) of items it contains are appropriate for the test design and adequately reflect the blueprint in terms of: <ul style="list-style-type: none"> ○ Full range of the State’s grade-level academic content standards; ○ Balance of content; ○ Cognitive complexity for each standard tested; ○ Range of item difficulty levels for each standard tested; ○ Structure of the assessment (e.g., number of items and proportion of item and response types specified by the blueprints); ○ Item pool size and composition sufficient to avoid over-exposure of items; • Results of an alignment study confirming that the test forms generated for individual students are aligned to the State’s academic content standards in terms of: <ul style="list-style-type: none"> ○ Full range of the State’s grade-level academic content standards; ○ Balance of content; ○ Cognitive complexity for each standard tested; ○ Range of item difficulty levels for each standard tested; ○ Structure of the assessment (i.e., features specified in Critical Element 2.1 – Test Design and Development, such as number of items and proportion of item and response types specified by the blueprints); • Empirical analyses that show: <ul style="list-style-type: none"> ○ The actual test forms produce an adequately precise estimate of student achievement; ○ Students are appropriately routed to the next item or stage based on their responses to the previous item or stage; ○ Response data adequately fit the psychometric model selected by the State. <p>AA-AAAS. For the State’s AA-AAAS, evidence to document adequate validity evidence based on content includes:</p> <ul style="list-style-type: none"> • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: <ul style="list-style-type: none"> ○ Test blueprints and other evidence submitted under Critical Element 2.1 – Test Design and Development; ○ Evidence documenting adequate linkage between the assessments and the academic content they are intended to measure; ○ Other documentation that shows the State’s assessments measure only the knowledge and skills specified in the State’s academic content standards (or extended academic content standards, as applicable) for the tested grade (i.e., not unrelated content);
--	---



	<ul style="list-style-type: none"> • Evidence of alignment, such as: <ul style="list-style-type: none"> ○ Report of results of an independent alignment study that is technically sound and document adequate linkage between each of the State’s assessments and the academic content the assessments are designed to measure; ○ If the State developed extended academic content standards for students with the most significant cognitive disabilities and used these to develop its AA-AAAS, the alignment study should document the linkage between the State’s academic content standards and extended academic content standards as well as adequate linkage between the extended academic content standards and the assessments; • For an adaptive AA-AAAS: <ul style="list-style-type: none"> ○ Summary of an analysis to confirm that the item pool adequately represents the test blueprints, such as a crosswalk of the item pool and the test blueprints; ○ Results of an alignment study that confirm that the test design, as implemented, produces assessments with adequate linkage to the academic content standards the assessments are designed to measure.
--	--

Critical Element 3.2 – Validity Based on Cognitive Processes

	Examples of Evidence
<p>The State has documented adequate validity evidence that its assessments tap the intended cognitive processes appropriate for each grade level as represented in the State’s academic content standards.</p>	<p>Evidence to support this critical element for the State’s general assessments includes:</p> <ul style="list-style-type: none"> • Validity evidence based on cognitive processes that shows levels of validity generally consistent with expectations of current professional standards, such as: <ul style="list-style-type: none"> ○ Results of cognitive labs exploring student performance on items that show the items require complex demonstrations or applications of knowledge and skills; ○ Reports of expert judgment of items that show the items require complex demonstrations or applications of knowledge and skills; ○ Empirical evidence that shows the relationships of items intended to require complex demonstrations or applications of knowledge and skills to other measures that require similar levels of cognitive complexity in the content area (e.g., teacher ratings of student performance, student performance on performance tasks or external assessments of the same knowledge and skills). <p>AA-AAAS. For the State’s AA-AAAS, evidence to support this critical element includes:</p> <ul style="list-style-type: none"> • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: <ul style="list-style-type: none"> ○ Results of cognitive labs exploring student performance on items that show the items require demonstrations or applications of knowledge and skills; ○ Reports of expert judgment of items that show the items require demonstrations or applications of knowledge and skills; ○ Empirical evidence that shows the relationships of items intended to require demonstrations or applications of knowledge and skills to other measures that require similar levels of cognitive complexity in the content



	area (e.g., teacher ratings of student performance, student performance on performance tasks or external assessments of the same knowledge and skills).
--	---

Critical Element 3.3 – Validity Based on Internal Structure

	Examples of Evidence
<p>The State has documented adequate validity evidence that the scoring and reporting structures of its assessments are consistent with the sub-domain structures of the State’s academic content standards on which the intended interpretations and uses of results are based.</p>	<p>Evidence to support this critical element for the State’s general assessments includes:</p> <ul style="list-style-type: none"> • Validity evidence based on the internal structure of the assessments that shows levels of validity generally consistent with expectations of current professional standards, such as: <ul style="list-style-type: none"> ○ Reports of analyses of the internal structure of the assessments (e.g., tables of item correlations) that show the extent to which the interrelationships among subscores are consistent with the State’s academic content standards for relevant student groups; ○ Reports of analyses that show the dimensionality of the assessment is consistent with the structure of the State’s academic content standards and the intended interpretations of results; ○ Evidence that ancillary constructs needed for success on the assessments do not provide inappropriate barriers for measuring the achievement of all students, such as evidence from cognitive labs or documentation of item development procedures; ○ Reports of differential item functioning (DIF) analyses that show whether particular items (e.g., essays, performance tasks, or items requiring specific knowledge or skills) function differently for relevant student groups. <p>AA-AAAS. For the State’s AA-AAAS, evidence to support this critical element includes:</p> <ul style="list-style-type: none"> • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: <ul style="list-style-type: none"> ○ Validity evidence based on the internal structure of the assessments, such as analysis of response patterns for administered items (e.g., student responses indicating no attempts at answering questions or suggesting guessing).



Critical Element 3.4 – Validity Based on Relations to Other Variables

	Examples of Evidence
<p>The State has documented adequate validity evidence that the State’s assessment scores are related as expected with other variables.</p>	<p>Evidence to support this critical element for the State’s general assessments includes:</p> <ul style="list-style-type: none"> • Validity evidence that shows the State’s assessment scores are related as expected with criterion and other variables for all student groups, such as: <ul style="list-style-type: none"> ○ Reports of analyses that demonstrate positive correlations between State assessment results and external measures that assess similar constructs, such as NAEP, TIMSS, assessments of the same content area administered by some or all districts in the State, and college-readiness assessments; ○ Reports of analyses that demonstrate convergent relationships between State assessment results and measures other than test scores, such as performance criteria, including college- and career-readiness (e.g., college-enrollment rates; success in related entry-level, college credit-bearing courses; post-secondary employment in jobs that pay living wages); ○ Reports of analyses that demonstrate positive correlations between State assessment results and other variables, such as academic characteristic of test takers (e.g., average weekly hours spent on homework, number of advanced courses taken); ○ Reports of analyses that show stronger positive relationships with measures of the same construct than with measures of different constructs; ○ Reports of analyses that show assessment scores at tested grades are positively correlated with teacher judgments of student readiness at entry in the next grade level. <p>AA-AAAS. For the State’s AA-AAAS, evidence to support this critical element includes:</p> <ul style="list-style-type: none"> • Validity evidence that shows levels of validity generally considered adequate by professional judgment regarding such assessments, such as: <ul style="list-style-type: none"> ○ Validity evidence based on relationships with other variables, such as analyses that demonstrate positive correlations between assessment results and other variables, for example: <ul style="list-style-type: none"> ▪ Correlations between assessment results and variables related to test takers (e.g., instructional time on content based on grade-level content standards); ▪ Correlations between proficiency on the high-school assessments and performance in post-secondary education, vocational training or employment.



SECTION 4: TECHNICAL QUALITY – OTHER

Critical Element 4.1 – Reliability

	Examples of Evidence
<p>The State has documented adequate reliability evidence for its assessments for the following measures of reliability for the State’s student population overall and each student group and, if the State’s assessments are implemented in multiple States, for the assessment overall and each student group, including:</p> <ul style="list-style-type: none"> • Test reliability of the State’s assessments estimated for its student population; • Overall and conditional standard error of measurement of the State’s assessments; • Consistency and accuracy of estimates in categorical classification decisions for the cut scores and achievement levels based on the assessment results; • For computer-adaptive tests, evidence that the assessments produce test forms with adequately precise estimates of a student’s achievement. 	<p>Collectively, evidence for the State’s general assessments and AA-AAAS must document adequate reliability evidence generally consistent with expectations of current professional standards.</p> <p>Evidence to support this critical element for the State’s general assessments includes documentation such as:</p> <ul style="list-style-type: none"> • A chapter on reliability in the technical report for the State’s assessments that shows reliability evidence; • For the State’s general assessments, documentation of reliability evidence generally consistent with expectations of current professional standards, including: <ul style="list-style-type: none"> ○ Results of analyses for alternate-form or, test-retest internal consistency reliability statistics, as appropriate, for each assessment; ○ Report of standard errors of measurement and conditional standard errors of measurement, for example, in terms of one or more coefficients or IRT-based test information functions at each cut score specified in the State’s academic achievement standards; ○ Results of estimates of decision consistency and accuracy for the categorical decisions (e.g., classification of proficiency levels) based on the results of the assessments. <p> For the State’s computer-adaptive assessments, evidence that estimates of student achievement are adequately precise includes documentation such as:</p> <ul style="list-style-type: none"> • Summary of empirical analyses showing that the estimates of student achievement are adequately precise for the intended interpretations and uses of the student’s assessment score; • Summary of analyses that demonstrates that the test forms are adequately precise across all levels of ability in the student population overall and for each student group (e.g., analyses of the test information functions and conditional standard errors of measurement). <p>AA-AAAS. For the State’s AA-AAAS, evidence to support this critical element includes:</p> <ul style="list-style-type: none"> • Reliability evidence that shows levels of reliability generally considered adequate by professional judgment regarding such assessments includes documentation such as: <ul style="list-style-type: none"> ○ Internal consistency coefficients that show that item scores are related to a student’s overall score; ○ Correlations of item responses to student proficiency level classifications; ○ Generalizability evidence such as evidence of fidelity of administration; ○ As appropriate and feasible given the size of the tested population, other reliability evidence as outlined above.



Critical Element 4.2 – Fairness and Accessibility

	Examples of Evidence
<p>The State has taken reasonable and appropriate steps to ensure that its assessments are accessible to all students and fair across student groups in the design, development and analysis of its assessments.</p>	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes:</p> <p>For the State’s general assessments:</p> <ul style="list-style-type: none"> • Documentation of steps the State has taken in the design and development of its assessments, such as: <ul style="list-style-type: none"> ○ Documentation describing approaches used in the design and development of the State’s assessments (e.g., principles of universal design, language simplification, accessibility tools and features embedded in test items or available as an accompaniment to the items); ○ Documentation of the approaches used for developing items; ○ Documentation of procedures used for maximizing accessibility of items during the development process, such as guidelines for accessibility and accessibility tools and features included in item specifications; ○ Description or examples of instructions provided to item writers and reviewers that address writing accessible items, available accessibility tools and features, and reviewing items for accessibility; ○ Documentation of procedures for developing and reviewing items in alternative formats or substitute items and for ensuring these items conform with item specifications; ○ Documentation of routine bias and sensitivity training for item writers and reviewers; ○ Documentation that experts in the assessment of students with disabilities, English learners and individuals familiar with the needs of other student populations in the State were involved in item development and review; ○ Descriptions of the processes used to write, review, and evaluate items for bias and sensitivity; ○ Description of processes to evaluate items for bias during pilot and field testing; ○ Evidence submitted under Critical Elements 2.1 – Test Design and Development and Critical Element 2.2 – Item Development; • Documentation of steps the State has taken in the analysis of its assessments, such as results of empirical analyses (e.g., DIF and differential test functioning (DTF) analyses) that identify possible bias or inconsistent interpretations of results across student groups. <p>AA-AAAS. For the State’s AA-AAAS:</p> <ul style="list-style-type: none"> • Documentation of steps the State has taken in the design and development of its assessments, as listed above; • Documentation of steps the State has taken in the analysis of its assessments, for example: <ul style="list-style-type: none"> ○ Results of bias reviews or, when feasible given the size of the tested student population, empirical analyses (e.g., DIF analyses and DTF analyses by disability category); ○ Frequency distributions of the tested population by disability category; ○ As appropriate, applicable and feasible given the size of the tested population, other evidence as outlined above. <p>Note: This critical element is closely related to Critical Element 2.2 – Item Development.</p>



Critical Element 4.3 – Full Performance Continuum

	Examples of Evidence
<p>The State has ensured that each assessment provides an adequately precise estimate of student performance across the full performance continuum, including for high- and low-achieving students.</p>	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes documents such as:</p> <p>For the State’s general assessments:</p> <ul style="list-style-type: none"> • Description of the distribution of cognitive complexity and item difficulty indices that demonstrate the items included in each assessment adequately cover the full performance continuum; • Analysis of test information functions (TIF) and ability estimates for students at different performance levels across the full performance continuum or a pool information function across the full performance continuum; • Table of conditional standard errors of measurement at various points along the score range. <p>AA-AAAS. For the State’s AA-AAAS:</p> <ul style="list-style-type: none"> • A cumulative frequency distribution or histogram of student scores for each grade and subject on the most recent administration of the State’s assessment; • For students at the lowest end of the performance continuum (e.g., pre-symbolic language users or students with no consistent communicative competencies), evidence that the assessments provide appropriate performance information (e.g., communicative competence); • As appropriate, applicable and feasible given the size of the tested population, other evidence as outlined above.

Critical Element 4.4 – Scoring

	Examples of Evidence
<p>The State has established and documented standardized scoring procedures and protocols for its assessments that are designed to produce reliable results, facilitate valid score interpretations, and report assessment results in terms of the State’s academic achievement standards.</p>	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes:</p> <ul style="list-style-type: none"> • A chapter on scoring in a technical report for the assessments or other documentation that describes scoring procedures, including: <ul style="list-style-type: none"> ○ Procedures for constructing scales used for reporting scores and the rationale for these procedures; ○ Scale, measurement error, and descriptions of test scores; • For scoring involving human judgment: <ul style="list-style-type: none"> ○ Evidence that the scoring of constructed-response items and performance tasks includes adequate procedures and criteria for ensuring and documenting inter-rater reliability (e.g., clear scoring rubrics, adequate training for and qualifying of raters, evaluation of inter-rater reliability, and documentation of quality control procedures); ○ Results of inter-rater reliability of scores on constructed-response items and performance tasks; • For machine scoring of constructed-response items: <ul style="list-style-type: none"> ○ Evidence that the scoring algorithm and procedures are appropriate, such as descriptions of development and calibration, validation procedures, monitoring, and quality control procedures;



	<ul style="list-style-type: none"> ○ Evidence that machine scoring produces scores are comparable to those produced by human scorers, such as rater agreement rates for human- and machine-scored samples of responses (e.g., by student characteristics such as varying achievement levels and student groups), systematic audits and rescoring; ● Documentation that the system produces student results in terms of the State’s academic achievement standards; ● Documentation that the State has rules for invalidating test results when necessary (e.g., non-attempt, cheating, unauthorized accommodation or modification) and appropriate procedures for implementing these rules (e.g., operations manual for the State’s assessment and accountability systems, test coordinators manuals and test administrator manuals, or technical reports for the assessments). <p>AA-AAAS. For the State’s AA-AAAS, in addition to the above:</p> <ul style="list-style-type: none"> ● If the assessments are portfolio assessments, evidence of procedures to ensure that only student work including content linked to the State’s grade-level academic content standards is scored; ● If the alternate assessments involve any scoring of performance tasks by test administrators (e.g., teachers): <ul style="list-style-type: none"> ○ Evidence of adequate training for all test administrators (may include evidence submitted under Critical Element 2.3 – Test Administration); ○ Procedures the State uses for each test administration to ensure the reliability of scoring; ○ Documentation of the inter-rater reliability of scoring by test administrators.
--	---

Critical Element 4.5 – Multiple Assessment Forms

	Examples of Evidence
<p>If the State administers multiple forms within a content area and grade level, within or across school years, the State ensures that all forms adequately represent the State’s academic content standards and yield consistent score interpretations such that the forms are comparable within and across school years.</p>	<p>Evidence to support this critical element for the State’s assessments system includes documents such as:</p> <ul style="list-style-type: none"> ● Documentation of technically sound equating procedures and results within an academic year, such as a section of a technical report for the assessments that provides detailed technical information on the method used to establish linkages and on the accuracy of equating functions; ● As applicable, documentation of year-to-year equating procedures and results, such as a section of a technical report for the assessments that provides detailed technical information on the method used to establish linkages and on the accuracy of equating functions.

Critical Element 4.6 – Multiple Versions of an Assessment

	Examples of Evidence
<p>If the State administers assessments in multiple versions within a content area, grade level, or school year, the State:</p>	<p>Evidence to support this critical element for the State’s general and alternate assessments includes:</p> <p>For the State’s general assessments:</p>



<ul style="list-style-type: none"> Followed a design and development process to support comparable interpretations of results for students tested across the versions of the assessments; Documented adequate evidence of comparability of the meaning and interpretations of the assessment results. 	<ul style="list-style-type: none"> Documentation that the State followed a design and development process to support comparable interpretations of results across different versions of the assessments (e.g., technology-based and paper-based assessments, assessments in English and native language(s), general and alternate assessments based on grade-level academic achievement standards); <ul style="list-style-type: none"> For a native language assessment, this may include a description of the State’s procedures for translation or trans-adaptation of the assessment or a report of analysis of results of back-translation of a translated test; For technology-based and paper-based assessments, this may include demonstration that the provision of paper-based substitutes for technology-enabled items elicits comparable response processes and produces an adequately aligned assessment; Report of results of a comparability study of different versions of the assessments that is technically sound and documents evidence of comparability generally consistent with expectations of current professional standards. <p> If the State administers technology-based assessments that are delivered by different types of devices (e.g., desktop computers, laptops, tablets), evidence includes:</p> <ul style="list-style-type: none"> Documentation that test-administration hardware and software (e.g., screen resolution, interface, input devices) are standardized across unaccommodated administrations; or Either: <ul style="list-style-type: none"> Reports of research (quantitative or qualitative) that show that variations resulting from different types of delivery devices do not alter the interpretations of results; or A comparability study, as described above. <p>AA-AAAS. For the State’s AA-AAAS:</p> <ul style="list-style-type: none"> Documentation that the State followed design, development and test administration procedures to ensure comparable results across different versions of the assessments, such as a description of the processes in the technical report for the assessments or a separate report.
---	--

Critical Element 4.7 – Technical Analysis and Ongoing Maintenance

	Examples of Evidence
<p>The State has a system for monitoring and 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 (i.e., general assessments and alternate assessments).</p>	<p>Evidence to support this critical element for the State’s assessments system includes:</p> <ul style="list-style-type: none"> Documentation that the State has established and implemented clear and technically sound criteria for analyses of its assessment system, such as: <ul style="list-style-type: none"> Sections from the State’s assessment contract that specify the State’s expectations for analyses to provide evidence of validity, reliability, and fairness; for independent studies of alignment and comparability, as appropriate; and for requirements for technical reports for the assessments and the content of such reports applicable to each administration of the assessment; The most recent technical reports for the State’s assessments that present technical analyses of the State’s



	<p>assessments;</p> <ul style="list-style-type: none"> ○ Documentation of the alignment of the State’s assessments to the State’s academic content standards (e.g., evidence submitted under Critical Element 3.1 – Overall Validity, Including Validity Based on Content; ○ Presentations of assessments results (e.g., to the State’s TAC); ● Documentation of the State’s system for monitoring and improving, as needed, the on-going quality of its assessment system, such as: <ul style="list-style-type: none"> ○ Evidence that the State has established and implemented clear criteria for the analysis of its assessment system (see above); ○ Documentation of regular internal and external technical review of components of the State’s assessment system, such as State Board of Education minutes, minutes from TAC meetings, and documentation of roles and responsibilities of TAC members; ○ Outline of a deliberate cycle for reviewing and updating the State’s academic content standards and assessments (e.g., provides for logical transitions such that the assessments are aligned to the standards on which instruction is based in the relevant school year).
--	--



SECTION 5: INCLUSION OF ALL STUDENTS

Critical Element 5.1 – Procedures for Including Students with Disabilities

	Examples of Evidence
<p>The State has in place procedures to ensure the inclusion of all public elementary and secondary school students with disabilities in the State’s assessment system, including, at a minimum, guidance for IEP Teams to inform decisions about student assessments that:</p> <ul style="list-style-type: none"> • Provides clear explanations of the differences between assessments based on grade-level academic achievement standards and assessments based on alternate academic achievement standards, including any effects of State and local policies on a student’s education resulting from taking an alternate assessment based on alternate academic achievement standards; • States that decisions about how to assess students with disabilities must be made by a student’s IEP Team based on each student’s individual needs; • Provides guidelines for determining whether to assess a student on the general assessment without accommodation(s), the general assessment with accommodation(s), or an alternate assessment; • Provides information on accessibility tools and features available to students in general and assessment accommodations available for students with disabilities; 	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • Documentation that the State has in place procedures to ensure the inclusion of all students with disabilities, such as: <ul style="list-style-type: none"> ○ Guidance for IEP Teams and IEP templates for students in tested grades; ○ Training materials for IEP Teams; ○ Accommodations manuals or other key documents that provide information on accommodations for students with disabilities; ○ Test administration manuals or other key documents that provide information on available accessibility tools and features; • Documentation that the implementation of the State’s alternate academic achievement standards promotes student access to the general curriculum, such as: <ul style="list-style-type: none"> ○ State policies that require that instruction for students with the most significant cognitive disabilities be linked to the State’s grade-level academic content standards; ○ State policies that require standards-based IEPs linked to the State’s grade-level academic content standards for students with the most significant cognitive disabilities; ○ Reports of State monitoring of IEPs that document the implementation of IEPs linked to the State’s grade-level academic content standards for students with the most significant cognitive disabilities. <p>Note: Key topics related to the assessment of students with disabilities are also addressed in Critical Element 4.2 -- Fairness and Accessibility and in critical elements addressing the AA-AAAS throughout.</p>



<ul style="list-style-type: none"> • Provides guidance regarding selection of appropriate accommodations for students with disabilities; • Includes instructions that students eligible to be assessed based on alternate academic achievement standards may be from any of the disability categories listed in the IDEA; • Ensures that parents of students with the most significant cognitive disabilities are informed that their student’s achievement will be based on alternate academic achievement standards and of any possible consequences of taking the alternate assessments resulting from district or State policy (e.g., ineligibility for a regular high school diploma if the student does not demonstrate proficiency in the content area on the State’s general assessments); • The State has procedures in place to ensure that its implementation of alternate academic achievement standards for students with the most significant cognitive disabilities promotes student access to the general curriculum. 	
---	--

Critical Element 5.2 – Procedures for Including English Learners

	Examples of Evidence
<p>The State has in place procedures to ensure the inclusion of all English learners in public elementary and secondary schools in the State’s assessment system and clearly</p>	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • Documentation of procedures for determining student eligibility for accommodations and guidance on selection of appropriate accommodations for English learners; • Accommodations manuals or other key documents that provide information on accommodations for English learners;



<p>communicates this information to districts, schools, teachers, and parents, including, at a minimum:</p> <ul style="list-style-type: none"> • Procedures for determining whether an English learner should be assessed with accommodation(s); • Information on accessibility tools and features available to all students and assessment accommodations available for English learners; • Guidance regarding selection of appropriate accommodations for English learners. 	<ul style="list-style-type: none"> • Test administration manuals or other key documents that provide information on available accessibility tools and features; • Guidance in key documents that indicates all accommodation decisions must be based on individual student needs and provides suggestions regarding what types of accommodations may be most appropriate for students with various levels of proficiency in their first language and English. <p>Note: Key topics related to the assessment of English learners are also addressed in Critical Element 4.2 – Fairness and Accessibility.</p>
--	--

Critical Element 5.3 – Accommodations

	Examples of Evidence
<p>The State makes available appropriate accommodations and ensures that its assessments are accessible to students with disabilities and English learners. Specifically, the State:</p> <ul style="list-style-type: none"> • Ensures that appropriate accommodations are available for students with disabilities under IDEA and students covered by Section 504; • Ensures that appropriate accommodations are available for English learners; • Has determined that the accommodations it provides (i) are appropriate and effective for meeting the individual student’s need(s) to participate in the assessments, (ii) do not alter the construct being assessed, and (iii) allow meaningful interpretations of results and comparison of scores for students who need and receive 	<p>Evidence to support this critical element for both the State’s general and AA-AAAS includes:</p> <ul style="list-style-type: none"> • Lists of accommodations available for students with disabilities under IDEA, students covered by Section 504 and English learners that are appropriate and effective for addressing barrier(s) faced by individual students (i.e., disability and/or language barriers) and appropriate for the assessment mode (e.g., paper-based vs. technology-based), such as lists of types of available accommodations in an accommodations manual, test coordinators manual or test administrators manual; • Documentation that scores for students based on assessments administered with allowable accommodations (and accessibility tools and features, as applicable) allow for valid inferences, such as: <ul style="list-style-type: none"> ○ Description of the reasonable and appropriate basis for the set of accommodations offered on the assessments, such as a literature review, empirical research, recommendations by advocacy and professional organizations, and/or consultations with the State’s TAC, as documented in a section on test design and development in the technical report for the assessments; ○ For accommodations not commonly used in large-scale State assessments, not commonly used in the manner adopted for the State’s assessment system, or newly developed accommodations, reports of studies, data analyses, or other evidence that indicate that scores based on accommodated and non-accommodated administrations can be meaningfully compared; ○ A summary of the frequency of use of each accommodation on the State’s assessments by student characteristics (e.g., students with disabilities, English learners); • Evidence that the State has a process to review and approve requests for assessment accommodations beyond those routinely allowed, such as documentation of the State’s process as communicated to district and school test coordinators and test administrators.



<p>accommodations and students who do not need and do not receive accommodations;</p> <ul style="list-style-type: none"> • Has a process to individually review and allow exceptional requests for a small number of students who require accommodations beyond those routinely allowed. 	
---	--

Critical Element 5.4 – Monitoring Test Administration for Special Populations

	Examples of Evidence
<p>The State monitors test administration in its districts and schools to ensure that appropriate assessments, with or without appropriate accommodations, are selected for students with disabilities under IDEA, students covered by Section 504, and English learners so that they are appropriately included in assessments and receive accommodations that are:</p> <ul style="list-style-type: none"> • Consistent with the State’s policies for accommodations; • Appropriate for addressing a student’s disability or language needs for each assessment administered; • Consistent with accommodations provided to the students during instruction and/or practice; • Consistent with the assessment accommodations identified by a student’s IEP Team or 504 team for students with disabilities, or another process for an English learner; • Administered with fidelity to test administration procedures. 	<p>Evidence to support this critical element for the State’s assessment system includes documents such as:</p> <ul style="list-style-type: none"> • Description of procedures the State uses to monitor that accommodations selected for students with disabilities, students covered by Section 504, and English learners are appropriate; • Description of procedures the State uses to monitor that students with disabilities are placed by IEP Teams in the appropriate assessment; • The State’s written procedures for monitoring the use of accommodations during test administration, such as guidance provided to districts; instructions and protocols for State, district and school staff; and schedules for monitoring; • Summary of results of monitoring for the most recent year of test administration in the State.



SECTION 6: ACADEMIC ACHIEVEMENT STANDARDS AND REPORTING

Critical Element 6.1 – State Adoption of Academic Achievement Standards for All Students

	Examples of Evidence
<p>The State formally adopted challenging academic achievement standards in reading/language arts, mathematics and in science for all students, specifically:</p> <ul style="list-style-type: none"> • The State formally adopted academic achievement standards in the required tested grades and, at its option, also alternate academic achievement standards for students with the most significant cognitive disabilities; • The State applies its grade-level academic achievement standards to all public elementary and secondary school students enrolled in the grade to which they apply, with the exception of students with the most significant cognitive disabilities to whom alternate academic achievement standards may apply; • The State’s academic achievement standards and, as applicable, alternate academic achievement standards, include: (a) At least three levels of achievement, with two for high achievement and a third for lower achievement; (b) descriptions of the competencies associated with each achievement level; and (c) achievement scores that differentiate among the achievement levels. 	<p>Evidence to support this critical element for the State’s assessment system includes:</p> <ul style="list-style-type: none"> • Evidence of adoption of the State’s academic achievement standards and, as applicable, alternate academic achievement standards, in the required tested grades and subjects (i.e., in reading/language arts and mathematics for each of grades 3-8 and high school and in science for each of three grade spans (3-5, 6-9, and 10-12)), such as State Board of Education minutes, memo announcing formal approval from the Chief State School Officer to districts, legislation, regulations, or other binding approval of academic achievement standards and, as applicable, alternate academic achievement standards; • State statutes, regulations, policy memos, State Board of Education minutes, memo from the Chief State School Officer to districts or other key documents that clearly state that the State’s academic achievement standards apply to all public elementary and secondary school students in the State (with the exception of students with the most significant cognitive disabilities to whom alternate academic achievement standards may apply); • Evidence regarding the academic achievement standards and, as applicable, alternate academic achievement standards, regarding: (a) at least three levels of achievement, including two levels of high achievement (e.g., proficient and advanced) and a third of lower achievement (e.g., basic); (b) descriptions of the competencies associated with each achievement level; and (c) achievement scores (i.e., “cut scores”) that differentiate among the achievement levels.



Critical Element 6.2 – Achievement Standards Setting

	Examples of Evidence
<p>The State used a technically sound method and process that involved panelists with appropriate experience and expertise for setting its academic achievement standards and alternate academic achievement standards to ensure they are valid and reliable.</p>	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes:</p> <ul style="list-style-type: none"> • The State’s standards-setting report, including: <ul style="list-style-type: none"> ○ A description of the standards-setting method and process used by the State; ○ The rationale for the method selected; ○ Documentation that the method used for setting cut scores allowed panelists to apply their knowledge and experience in a reasonable manner and supported the establishment of reasonable and defensible cut scores; ○ Documentation of the process used for setting cut scores and developing performance-level descriptors aligned to the State’s academic content standards; ○ A description of the process for selecting panelists; ○ Documentation that the standards-setting panels consisted of panelists with appropriate experience and expertise, including: <ul style="list-style-type: none"> ▪ Content experts with experience teaching the State’s academic content standards in the tested grades; ▪ Individuals with experience and expertise teaching students with disabilities, English learners and other student populations in the State; ▪ As appropriate, individuals from institutions of higher education (IHE) and individuals knowledgeable about career-readiness; ▪ A description, by relevant characteristics, of the panelists (overall and by individual panels) who participated in achievement standards setting; ○ If available, a summary of statistical descriptions and analyses that provides evidence of the reliability of the cut scores and the validity of recommended interpretations. • AA-AAAS. For the State’s AA-AAAS, in addition to the above: <ul style="list-style-type: none"> • Documentation that the panels for setting alternate academic achievement standards included individuals knowledgeable about the State’s academic content standards and special educators knowledgeable about students with the most significant cognitive disabilities.

Critical Element 6.3 – Challenging and Aligned Academic Achievement Standards

	Examples of Evidence
<p>The State’s academic achievement standards are challenging and aligned with the State’s academic content standards such that a high school student who scores at the proficient or above level has mastered what students are expected to know and be able to do by the time</p>	<p>Evidence to support this critical element for the State’s general assessments and AA-AAAS includes:</p> <p>For the State’s general assessments:</p> <ul style="list-style-type: none"> • Documentation that the State’s academic achievement standards are aligned with the State’s academic content standards, such as: <ul style="list-style-type: none"> ○ A description of the process used to develop the State’s academic achievement standards that shows that: <ul style="list-style-type: none"> ▪ The State’s grade-level academic content standards were used as a main reference in writing



<p>they graduate from high school in order to succeed in college and the workforce.</p> <p>If the State has defined alternate academic achievement standards for students with the most significant cognitive disabilities, the alternate academic achievement standards are linked to the State’s grade-level academic content standards or extended academic content standards, show linkage to different content across grades, and reflect professional judgment of the highest achievement standards possible for students with the most significant cognitive disabilities.</p>	<p>performance level descriptors;</p> <ul style="list-style-type: none"> ▪ The process of setting cut scores used, as a main reference, performance level descriptors that reflect the State’s grade-level academic content standards; ▪ The State’s cut scores were set and performance level descriptors written to reflect the full range of the State’s academic content standards for each grade; ○ A description of steps taken to vertically articulate the performance level descriptors across grades; ○ Evaluation by standard-setting panelists or external expert reviewers that the State’s academic achievement standards are aligned to the grade-level academic content standards and include subject-specific performance level descriptors that meaningfully differentiate across performance levels within grades and are vertically articulated across grades; <ul style="list-style-type: none"> • Documentation that the State’s academic achievement standards are challenging, such as: <ul style="list-style-type: none"> ○ Reports of the results of benchmarking the State’s academic achievement standards against NAEP, international assessments or other related and appropriate measures; ○ Policies of the State network of institutions of higher education (IHEs) that exempt from remedial courses and place into credit-bearing college courses any student who scores at the proficient level or above on the State’s high school assessments. <p>AA-AAAS. For the State’s AA-AAAS:</p> <ul style="list-style-type: none"> • Documentation that the State’s alternate academic achievement standards are linked to the State’s academic content standards, such as: <ul style="list-style-type: none"> ○ A description of the process used to develop the alternate academic achievement standards that shows: <ul style="list-style-type: none"> ▪ The State’s grade-level academic content standards or extended academic content standards were used as a main reference in writing performance level descriptors for the alternate academic achievement standards; ▪ The process of setting cut scores used, as a main reference, performance level descriptors linked to the State’s grade-level academic content standards or extended academic content standards; ▪ The cut scores were set and performance level descriptors written to link to the State’s grade-level academic content standards or extended academic content standards; ▪ A description of steps taken to vertically articulate the alternate academic achievement standards (including cut scores and performance level descriptors) across grades.
---	--

Critical Element 6.4 – Reporting

	Examples of Evidence
<p>The State reports its assessment results, and the reporting facilitates timely, appropriate, credible, and defensible interpretations and uses of results for students tested by parents, educators,</p>	<p>Collectively, for the State’s assessment system, evidence to support this critical element must demonstrate that the State’s reporting system facilitates timely, appropriate, credible, and defensible interpretation and use of its assessment results.</p> <p>Evidence to support this critical element both the State’s general assessments and AA-AAAS includes:</p>



<p>State officials, policymakers and other stakeholders, and the public, including:</p> <ul style="list-style-type: none"> • The State reports to the public its assessment results on student achievement at each proficiency level and the percentage of students not tested for all students and each student group after each test administration; • The State reports assessment results, including itemized score analyses, to districts and schools so that parents, teachers, principals, and administrators can interpret the results and address the specific academic needs of students, and the State also provides interpretive guides to support appropriate uses of the assessment results; • The State provides for the production and delivery of individual student interpretive, descriptive, and diagnostic reports after each administration of its assessments that: <ul style="list-style-type: none"> ○ Provide valid and reliable information regarding a student’s achievement; ○ Report the student’s achievement in terms of the State’s grade-level academic achievement standards (including performance-level descriptors); ○ Provide information to help parents, teachers, and principals interpret the test results and address the specific academic needs of students; ○ Are available in alternate formats (e.g., Braille or large print) upon request and, to the extent 	<ul style="list-style-type: none"> • Evidence that the State reports to the public its assessment results on student achievement at each proficiency level and the percentage of students not tested for all students and each student group after each test administration, such as: <ul style="list-style-type: none"> ○ State report(s) of assessment results; ○ Appropriate interpretive guidance provided in or with the State report(s) that addresses appropriate uses and limitations of the data (e.g., when comparisons across student groups of different sizes are and are not appropriate). • Evidence that the State reports results for use in instruction, such as: <ul style="list-style-type: none"> ○ Instructions for districts, schools, and teachers for access to assessment results, such as an electronic database of results; ○ Examples of reports of assessment results at the classroom, school, district and State levels provided to teachers, principals, and administrators that include itemized score analyses, results according to proficiency levels, performance level descriptors, and, as appropriate, other analyses that go beyond the total score (e.g., analysis of results by strand); ○ Instructions for teachers, principals and administrators on the appropriate interpretations and uses of results for students tested that include: the purpose and content of the assessments; guidance for interpreting the results; appropriate uses and limitations of the data; and information to allow use of the assessment results appropriately for addressing the specific academic needs of students, student groups, schools and districts. ○ Timeline that shows results are reported to districts, schools, and teachers in time to allow for the use of the results in planning for the following school year. • Evidence to support this critical element for both general assessments and AA-AAAS, such as: <ul style="list-style-type: none"> ○ Templates or sample individual student reports for each content area and grade level for reporting student performance that: <ul style="list-style-type: none"> ▪ Report on student achievement according to the domains and subdomains defined in the State’s academic content standards and the achievement levels for the student scores (though sub-scores should only be reported when they are based on a sufficient number of items or score points to provide valid and reliable results); ▪ Report on the student’s achievement in terms of grade-level achievement using the State’s grade-level academic achievement standards and corresponding performance level descriptors; ▪ Display information in a uniform format and use simple language that is free of jargon and understandable to parents, teachers, and principals; ▪ Examples of the interpretive guidance that accompanies individual student reports, either integrated with the report or a separate page(s), including cautions related to the reliability of the reported scores; ▪ Samples of individual student reports in other languages and/or in alternative formats, as applicable. • Evidence that the State follows a process and timeline for delivering individual student reports, such as: <ul style="list-style-type: none"> ○ Timeline adhering to the need for the prompt release of assessment results that shows when individual
--	---



<p>practicable, in a native language that parents can understand;</p> <ul style="list-style-type: none"> • The State follows a process and timeline for delivering individual student reports to parents, teachers, and principals as soon as practicable after each test administration. 	<p>student reports are delivered to districts and schools;</p> <ul style="list-style-type: none"> ○ Key documents, such as a cover memo that accompanies individual student reports delivered to districts and schools, listserv messages to district and school test coordinators, or other meaningful communication to districts and schools that include the expectation that individual student reports be delivered to teachers and principals and corresponding expectations for timely delivery to parents (e.g., within 30 days of receipt). <p>Note: Samples of individual student reports and any other sample reports should be redacted to protect personally identifiable information, as appropriate, or populated with information about a fictitious student for illustrative purposes.</p>
--	--



A Technical Commentary on Arizona's Menu of Assessments Legislation (H.B. 2544)

Derek Briggs
University of Colorado
Jerry D'Agostino
Ohio State University

December 23, 2016



Overview

This white paper has two purposes. The first purpose is to discuss, from a technical standpoint, the feasibility of implementing Arizona’s H.B. 2544, which calls for the state board of education to adopt a “menu of locally procured achievement assessments to measure pupil achievement of the state academic standards.” This menu approach is intended to initially apply only to students in grades 9-12 as of the 2017-18 school year, but would then extend to students in grades 3-8 as well by the 2018-19 schools year. In discussing the feasibility of H.B. 2544, we focus attention on provision E.3, which stipulates, in essence, that the scores across the different assessments within the hypothetical menu should be interchangeable, such that regardless of which assessment a student is administered, the resulting inference about a student’s achievement level (i.e., proficiency, college readiness, etc.) should remain the same, and by extension, so should subsequent inferences about school-level performance. The second purpose of this paper is to provide specific recommendations with regard to the evidence that would be needed to evaluate not only provision E.3, but provisions E.1 and E.2.

Although we are sympathetic to the motivation behind H.B. 2544 (“the state should relieve students, teachers and schools of unnecessary duplicative testing and maximize instructional time”), we are very pessimistic that the law can be implemented in a manner that would meet the requirements as stipulated in E.1, E.2, and E.3. To the extent that scores from the menu of assessments would continue to serve as inputs for the state’s system of educational accountability, they would be very hard to defend from claims that certain schools are advantaged/disadvantaged by a strategic choice to emphasize one assessment option over another. Instead, we would recommend a much more limited version of the menu approach for 11th grade testing that relaxes the need for the scores from the menu of assessments to be used evidence of the status and growth of student achievement.

A Short Primer on Relevant Terminology from Educational Measurement

In order to discuss technical aspects necessary to meet the stipulations H. B. 2544 we invoke the sometimes esoteric terminology of educational measurement. Key terms include the *construct of measurement, linking, equating, scale aligning, and prediction*. To avoid confusion, we use this section to define these terms before using them in a subsequent section to point out obstacles that stand in the way of implementing H.B 2544.

Construct of Measurement

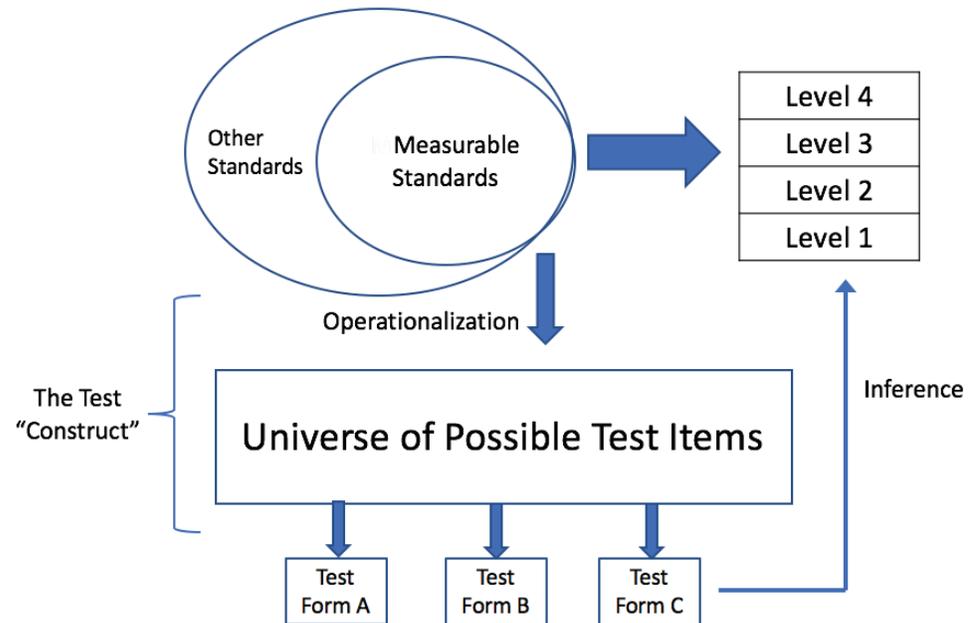


Figure 1. *Test Design and the Operationalization of a Construct of Measurement for a Given Grade/Course and Subject Domain*

In an educational achievement testing content, the *construct of measurement* represents a composite of the knowledge, skills and abilities students are expected to develop as they are formally exposed to curricula and instruction in school. As such, a construct is something that must be negotiated and operationalized before it can be “measured” in the form of a score on some collection of tests items. Figure 1 depicts this process of negotiation and operationalization that underlies any state assessment program. The starting point is typically a document that specifies, with varying degrees of specificity, standards for what students should know and be able to do in some academic subject at the culmination of a grade or course. Note that decisions must be made about the subset of these standards that can be plausibly measured within the constraints of a time-delimited standardized test. These measurable standards must then be organized and prioritized to form a basis for the differentiation of students in terms of achievement levels.

Together then, the measurable standards (and the achievement levels that are defined as a function of these standards) form the basis for a blueprint for item development. This blueprint specifies the breadth and depth of content coverage, necessary differences in item formats, rules for scoring student responses, and the way that items are to be selected for

inclusion on a test form. Importantly, the blueprint produces a pool of test items that are actually created at one point in time, but also can be seen as the basis for hypothetical test items that *could* (and eventually will) be created in the future. Because of this, every test blueprint is associated with a “universe” of possible test items. **It follows that the construct of measurement is the score we would observe if it were possible for a student to be given a test comprised of the complete universe of test items that derive from the test blueprint.** Because this is not possible, test forms comprised of item subsets are “sampled” to be representative of the universe of items. The scores from this sample of items are used to make inferences about a student’s proficiency with respect to the unknown construct. For any specific grade/course and subject, multiple test forms (e.g., A, B, C in Figure 1) are created for administration both within a given year and across years. So long as these forms have been assembled according to the same blueprint specifications, the scores from each form provide equally valid estimates of the construct of measurement—hence they can be **equated** to one another.

Given the definition above, if two different tests have been written according to two different blueprints, then strictly speaking, the tests provide measures of two different constructs. Questions about the comparability of scores from different tests hinge upon the degree of conceptual overlap that exists between the two constructs. When the overlap is large, it may be possible to link the scores together; when the overlap is small, the best that can be done is to predict one score from the other. Only when the overlap is complete is it possible to fully equate scores such that one can be treated as interchangeable for the other.

A Taxonomy for Linking Test Scores

Imagine we have two different tests, X and Y. Following Holland & Dorans (2006), we use the term *linking* to refer to the general idea of a transformation between the scores of one test and those of another. The linking methods that can be used for this transformation fall into three categories: equating, scale aligning (i.e., “scaling”) and predicting. **Equating** is the strongest linking method. When a link between tests has been established by equating scores on test X to scores on test Y, the scores can be considered interchangeable. Holland & Dorans (2006) outline five requirements viewed as necessary for the equating of test X and Y to be successful.

1. **The equal construct requirement.** The tests should measure the same construct.
2. **The equal reliability requirement.** The tests should have the same reliability.
3. **The symmetry requirement.** The equating function for equating the scores of Y to X should be the inverse of the function for equating the scores of X to Y.
4. **The equity requirement.** It should be a matter of indifference to an examinee to be tested by either one of two tests that have been equated (hence, they are interchangeable).
5. **The population invariance requirement.** The choice of (sub)population used to estimate the equating function between the scores of tests X and Y should not matter—the equating function should be population invariant.

Scaling is a weaker linking method that produces scores that are comparable, but not necessarily interchangeable, because at least one of the five criteria above is not met. In general, scaling methods have to rely upon stronger statistical assumptions, and these assumptions are at once seldom fully met and can also be difficult to evaluate empirically. **Predicting** is the weakest linking method. Although it is relatively easy to carry out and does allow for inferences to be made from the scores on one test to the scores on another, it will always violate, at a minimum, the symmetry requirement (i.e., two different test score conversion tables are needed, one that predicts scores on X from scores on Y, and another than predicts scores on Y from scores on X, where the latter function is NOT the inverse of the former).

Examples of Equating vs Scale Aligning vs. Predicting

- Two forms X and Y of the same test are created by the same vendor (or different vendors) according to the process depicted in Figure 1. Both test forms have the same number of items and are equally reliable. So long as certain design features are in place (see next subsection), it becomes a relatively straightforward and defensible task to **equate** scores on test form X to the scores on test form Y. In this example the first four of the Holland & Doran’s criteria can definitely be satisfied, and there is good reason to suspect that the fifth criterion (population invariance) will be met as well.
- Two tests X and Y are created by the same vendor for adjacent grades in the same subject. Each test has been created according to the process depicted in Figure 1, but on the basis of different sets of grade-level standards and hence two somewhat different blueprints for item design. The tests are equally reliable, take the same amount of time to administer, and are administered under the same standardized conditions. This is typical of the process used to create a vertically **aligned score scale** across grades. It hinges upon a design in which a common set of items are embedded in tests X and Y and administered to the different populations of students taking the two tests. A statistical model (i.e., a model from item response theory) is used to place the two tests onto a single comparable score scale. Although the linking in this example is likely to fail the equal construct criterion, it is still possible to attain some degree of comparability so long as the item response theory model can be shown to have adequate fit to the data.
- The SAT and ACT are created by two different vendors as tests of college readiness. It is assumed that there is some overlap in the process (i.e., Figure 1) used to create these tests, but the degree of overlap is unknown, nor do the vendors claim to be measuring the same construct. The tests differ somewhat in their reliability and take different amounts of time to administer. A self-selected subpopulation of students takes both tests. A statistical model (i.e., linear regression) is used to **predict** the scores on the SAT from the ACT and the predict the scores on the ACT from the SAT. In this example, the link that has been made between the two tests is unlikely to satisfy any of the five criteria for a successful equating of scores. We can predict the score a student would receive on one test given the score on the other, but this prediction will have a great

deal of uncertainty, and may not generalize to all students taking the two tests. Because the prediction will possess at least some degree of prediction error, one cannot claim the tests will yield interchangeable scores.

Data Collection Designs

Regardless of the method used to link test scores, a link can only be established when some design has been put in place to gather the necessary data. This must be done in a very purposeful manner and ideally should involve the random assignment of students to different tests. When students cannot be randomly assigned to either testing condition, the basis for linking tests becomes more equivocal, either depending upon a non-random group of students who take both tests, or upon the performance of all students on a non-random set of items common to either one or both of the tests. Importantly, because the links between different tests need to be constantly monitored and evaluated, data collection designs need to be maintained throughout, which would be time-consuming and costly.

Technical Problems with the Requirements of H.B. 2544

H.B. 2544 establishes three hurdles all locally procured assessments in the menu are expected to jump:

“E. The State Board of Education shall require that the provider of a locally procured achievement assessment that is proposed to be considered for the menu of locally procured achievement assessments shall

1. Provide evidence that the assessment is a high quality assessment.
2. Demonstrate that the assessment meets or exceeds the state board’s adopted academic standards.
3. Demonstrate that the assessment scores can be equated for state accountability programs including establishing comparable student assessment scores and performance levels for achievement profiles and letter grade classifications issued pursuant to section 15-241.”

All three of these stipulations are vaguely worded in ways that require clarification. Stipulations E.1 and E.2 are best subsumed within the at once global yet more specific statement “Provide evidence that the assessment is valid and reliable for its intended interpretation and uses.” This statement is consistent with language found in the *Standards for Educational and Psychological Testing* (2014), published jointly by the American Educational Research Association, the American Psychological Association and the National Council for Measurement in Education. In order for an assessment to be “high quality” there must be evidence of its validity and reliability; in order for an assessment to be valid in Arizona’s achievement testing context, it would be critical to demonstrate alignment with the state board’s academic standards.

Stipulation E.3 is problematic for two reasons. First, all three stipulations in E are intended to refer to evidence that a specific provider of a locally procured assessment is expected to produce, with the implication that this evidence would be necessary before an assessment would be approved for inclusion. However, while it would be possible for a vendor to “demonstrate” at least some evidence related to the validity of its assessment up front, it would not be possible to demonstrate that its scores lead to comparable student scores, performance levels and school classifications until a data collection design could be established and then analyzed in collaboration with ADE and other vendors of menu assessments. The language of E.3 seems to provide wiggle room in this regard by using more qualified language (“demonstrate that the assessment scores *can* be equated”). It is unclear whether the expectation is that a vendor would only need to provide evidence that they have a data collection design in place that would make it possible to evaluate comparability, or whether they would need to provide empirical evidence that comparability had already been established. Second, the (perhaps inadvertent) use of the term equated in E.3 establishes a hurdle that no provider will be able to demonstrate relative to the established standards and criteria (e.g., Holland & Dorans, 2006) for equating test scores. Although the constructs measured by assessments in the menu of assessments are sure to contain overlap, they are by definition (see Figure 1) not the same because the items on each test were written to satisfy different blueprints. Some assessments may have been written to provide evidence for standards that are not part of Arizona’s academic standards (in which case the scores from these assessments would contain what psychometricians refer to as “construct irrelevant variance”), or some assessments may not fully capture Arizona’s academic standards or otherwise diverge from the blueprint used to design the AzMERIT (in which case the scores from these menu assessments would suffer from what psychometricians refer to as “construct underrepresentation”). In either case, one would be hard-pressed to argue that the assessments measure identical constructs.

Furthermore, there is no guarantee that assessments in the menu will be equally reliable (another requirement for score equating). For example, the AzMERIT tests are developed with an eye toward measuring students with precision at all end of the achievement continuum, something that is especially important when test scores are being used to provide estimates of not just student achievement, but growth in student achievement. In contrast, the ACT and SAT are developed with an eye toward maximizing precision around the score that distinguishes students who are college ready from those who are not. Indeed, the entire notion that the assessments within a menu can be equated is on its face paradoxical. If tests have been equated, then students (and by extension parents) should be indifferent as to which test they take. But the whole point of H.B 2544 is that students (and parents) are not indifferent to this choice.

The near impossibility of arguing that assessments within the menu can be equated does not mean that it is impossible to establish weaker links through scale aligning techniques or score prediction techniques. These weaker links might at least facilitate comparisons between assessments that make it possible to evaluate, on average, whether students (and schools) are unfairly advantaged/disadvantaged by taking test X in place of test Y. However,

creating even these weaker links between test scores would require an ongoing (and costly) commitment by the state to data collection designs in which either random samples of students take multiple assessments from the menu, or (more plausibly) that a common short anchor test (i.e., “AzMERIT Lite”) would be given to all students as a supplement to their primary assessment choice from the menu. This presents its own unique set of challenges, and to date it is unprecedented for a state to attempt to simultaneously establish (and maintain) linking transformations not just from test X to test Y, but from test X (i.e., AzMERIT) to tests Q, R, S, and T (other hypothetical assessments on the menu). To accomplish this vendors for different assessments would need to collaborate with ADE (and ADE’s vendor for the AzMERIT, presently AIR) on joint data collection and analysis designs. Vendors would also have to establish a common approach to the way they maintain test security and provide students with testing accommodations. The logistical challenges involved in facilitating this sort of collaboration between vendors and ADE would be daunting.

The issues above were discussed during a two-day meeting of the state’s technical advisory committee (TAC) on November 7-8, 2016. The most constructive suggestion for a path forward was that if the state placed high value on flexibility and the need to reduce duplicative testing that this could be accomplished for grade 11 with the menu of assessments plan provided that the scores from these assessments would not be used for the achievement and growth components of the state’s accountability ratings for schools. Instead, the state could take advantage of ESSA’s flexibility to use *participation* in one assessment from the menu of assessments as part of a college readiness indicator. For high schools, grade 9 and 10 AzMERIT scores would continue to provide evidence relevant to achievement and growth for school-level accountability. If the use of test scores for purpose of school accountability were to be removed, then the issue of score comparability among assessments within an 11th grade menu would no longer be a predominant concern. It would be up to students and parents to decide on an assessment, and only they would assume the consequence of inadvertently choosing the “wrong” assessment.

Evidence Requirements for Stipulations E.1-E.3

Our recommendations above notwithstanding, for any vendor seeking to demonstrate the ability to fulfill stipulations E.1-E.3:

1. With respect to evidence for validity (i.e., E.1-E.2), vendors should use Chapter 1 from the AERA/APA/NCME’s *Standards for Educational and Psychological Testing* as a framework.
 - a. As Chapter 1 makes clear, a critical starting point is to establish the intended uses and interpretations for its test scores, and these intended uses and interpretations need to be specific to Arizona’s context.
 - b. Next, the vendor should establish whether the validity evidence it has already gathered can be reasonably generalized to Arizona’s educational context.
 - c. Finally, different forms of validity evidence should be organized with respect to (1) content-oriented evidence (in particular, evidence that the vendor’s items

and test blueprint are adequately aligned to the state's academic standards), (2) evidence regarding cognitive processes, (3) evidence regarding internal structure, (4) evidence regarding relationships with conceptually related constructs, (5) evidence regarding relationships with external criteria, and (6) evidence based on consequences of testing.

2. With respect to evidence for reliability (i.e., E.1), vendors should use Chapter 2 from the AERA/APA/NCME's *Standards for Educational and Psychological Testing* as a framework. In particular (e.g., Standard 2.14), vendors need to provide information not just about overall reliability estimates and the unconditional standard error of measurement, but also for conditional standard errors of measurement for students at different locations on their scale score continuum.
3. With respect to evidence of comparability (E.3), vendors should use Standards Cluster 3 (Standards 5.12-5.20) from Chapter 5 of AERA/APA/NCME's *Standards for Educational and Psychological Testing* as a framework.
 - a. Vendors should clarify whether (and on what grounds) they believe that scores from their assessment can be equated with the AzMERIT, made comparable via scale aligning techniques, or predicted from AzMERIT scores.
 - b. Vendors should provide concrete plans for analyses and criteria that they would enact to evaluate, on an ongoing basis, the population invariance of any proposed links between tests. A proposed data collection design should be part of these plans.
 - c. Vendors should specify steps they will take to ensure comparability with respect to test security, universal design and testing accommodations.

References

Holland, P. & Dorans, N. (2006). Linking and Equating. In R. Brennan (ed) *Educational Measurement*, 4th Edition.

AERA/APA/NCME (2014) *Standards for Educational and Psychological Testing*.
https://www.ncme.org/ncme/NCME/Publication/Testing_Standards/NCME/Publication/Testing_Standards.aspx?hkey=c5136771-5475-4ba9-8132-9bcc1ca5a277