The Wyoming Assessment Handbook

Updated

Spring 2008

Dr. Jim McBride, Wyoming Superintendent of Public Instruction

Wyoming Department of Education

Hathaway Building, 2nd Floor

2300 Capitol Ave

Cheyenne, WY 82002-0050

Acknowledgement

<u>Letter to Educators from Scott Marion and Sue Stevens</u>

March 15, 2001

Dear Wyoming Educator:

The standards-based reform movement has rapidly increased the pressure for educators to develop sophisticated assessment skills. We wrote this handbook to help provide much of the information necessary for you to design valid local assessment systems. This book is designed to be "user-friendly," but to also provide some in-depth discussion of complex assessment issues. Therefore, we expect most readers to have some familiarity with basic assessment principles.

This document would not have been possible without the help and collaboration of many individuals. First, thanks to the more than 90 Wyoming educators who attended the "Cody meeting" in June 2000, which helped formulate many of the ideas in this document. The ideas articulated in this handbook have been shaped by countless discussions with professionals at the Wyoming Department of Education, particularly Annette Bohling, Judy Catchpole, Mike Hamilton, Mary Kay Hill, Deb Holloway, Carol Mawford, Joe Simpson, and Becca Walk. Equally important were our discussions with many professionals working in Wyoming school districts especially, Craig Beck, Terry Bridwell, Teresa Chaulk, Matt Davidson, Bob Krisko, Rae Lynn Job, John Metcalfe, Pia Hansen Powell, Tom Sachse, Jim Staab, Teresa Staab, Judy Turner, and Ed Weber. Special thanks to Jim Lowham, Mike Flicek, and Connie Nerby for pushing our thinking and helping us come to a greater understanding of these issues. Several outside measurement professionals have provided considerable help in shaping our thinking and refining this document. Richard Hill, Mariam Manely, Alan Moore, and Phoebe Winter provided valuable comments on earlier drafts of the handbook. Ted Coladarci's careful editing and commentary and Dale Carlson's wise advice, critical questions, and encouragement were invaluable. Finally, Deb Holloway and Jeanie Hall critically read the final draft and their comments improved the final product. We are grateful for all of the help we have received and recognize that any credit must be shared widely. However, any errors or omissions are ours alone.

We hope you find this document useful and come to believe, as we do, that well-designed assessment systems can lead to the improvement of teaching and learning. We also hope that you share comments and criticisms with us so that we can try to improve subsequent versions of this handbook.

The Wyoming Department of Education does not discriminate on the basis of race, color, national origin, sex, age, or disability in admission or access to, or treatment or employment in its educational programs or activities. Inquiries concerning Title VI, Title IX, Section 504, and the Americans with Disabilities Act may be referred to the Wyoming Department of Education, Office for Civil Rights Coordinator, 2nd floor, Hathaway Building, Cheyenne, Wyoming 82002-0050 or (307) 777-2555, or the Office for Civil Rights, Region VIII, U. S. Department of Education, Federal Building, Suite 310, 1244 Speer Boulevard, Denver, CO 80204-3582, or (303) 844-5695 or TDD (303) 844-3417. This publication will be provided in an alternative format upon request.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION AND FRAMEWORK

School districts throughout the country are being bombarded with a multitude of assessment requirements. Based upon the assessment requirements in the No Child Left Behind Act, it does not appear that these assessment demands will diminish. In Wyoming, there are many assessment requirements for districts and schools. All of these requirements fit under the umbrella of the overall district assessment system. This overarching system will be used to satisfy Wyoming accreditation requirements. At the high school level, the "Body of Evidence" system will add more components at a finer level of detail to determine if students are eligible for graduation, while the state comprehensive assessment system carries a parallel function at the elementary level or other levels in districts in language arts, mathematics, and science. How can all these assessment requirements be met most efficiently? This document is designed to help districts develop a comprehensive assessment plan that can meet necessary requirements and promote student learning. This first chapter provides some background information about standards-based education followed by a discussion of the concept of an assessment system. The key principles necessary to keep in mind when designing assessment systems are presented in Chapter II. Chapter III contains a discussion of the district assessment system and the measurement requirements for Wyoming school district accreditation. The Body of Evidence Assessment System for graduation is presented in Chapter IV.

Standards-Based Education

During the latter part of the 20th and early 21st Century our society has undergone rapid changes that, in part, have led to demands to improve our educational system. In the early twentieth century, schools were given the task of sorting and selecting students. This system served society well, allowing students to find their place in a predominately agricultural and industrial economy. As the nation's economic system began to change, it became apparent that students needed higher level workplace and academic competencies. Our educational system is now expected to educate all students to a much higher level than in the past.

Almost all of the states are in the process of shifting to a standards-based reform model, a movement that can be traced to the 1983 publication of *A Nation at Risk*. A standards-based approach relies on a few key premises. Clear and meaningful learning targets (standards) are established, student progress toward meeting these targets is measured, and information from these assessments and other sources is used to improve student learning. This model relies heavily on information and responsibility. Everyone–students, parents, teachers, administrators, and policy makers–should know what students are expected to learn, how learning will be measured, and what the results imply for improving student learning. Everyone must also be responsible for fulfilling his or her role in the improvement effort. As indicated in Figure 1, this model includes critical elements necessary for a total education improvement system.

The centerpiece of a standards-based system is a set of challenging standards that hold high expectations for all students, regardless of students' backgrounds or where they attend school. The standards help everyone involved in the system focus on the academic performance of students indicated by the standards, rather than the resources or effort put into the system.

➢ **Content standards** define what students should know and be able to do as a result of instruction in the common core of knowledge and skill areas. These should be clear and rigorous.

➢ **Performance standards** describe the characteristics of students at various levels of performance. They describe "how good is good enough?" These should clearly differentiate what students "look like" at various stages of learning the content, and the performance standards should be clearly measurable.

➢ **Assessments** are designed to measure students' progress toward meeting the content standards at specific benchmarks.

➢ **School Delivery Standards** are designed to ensure that the system is providing adequate resources for all students so that they might meet the standards.
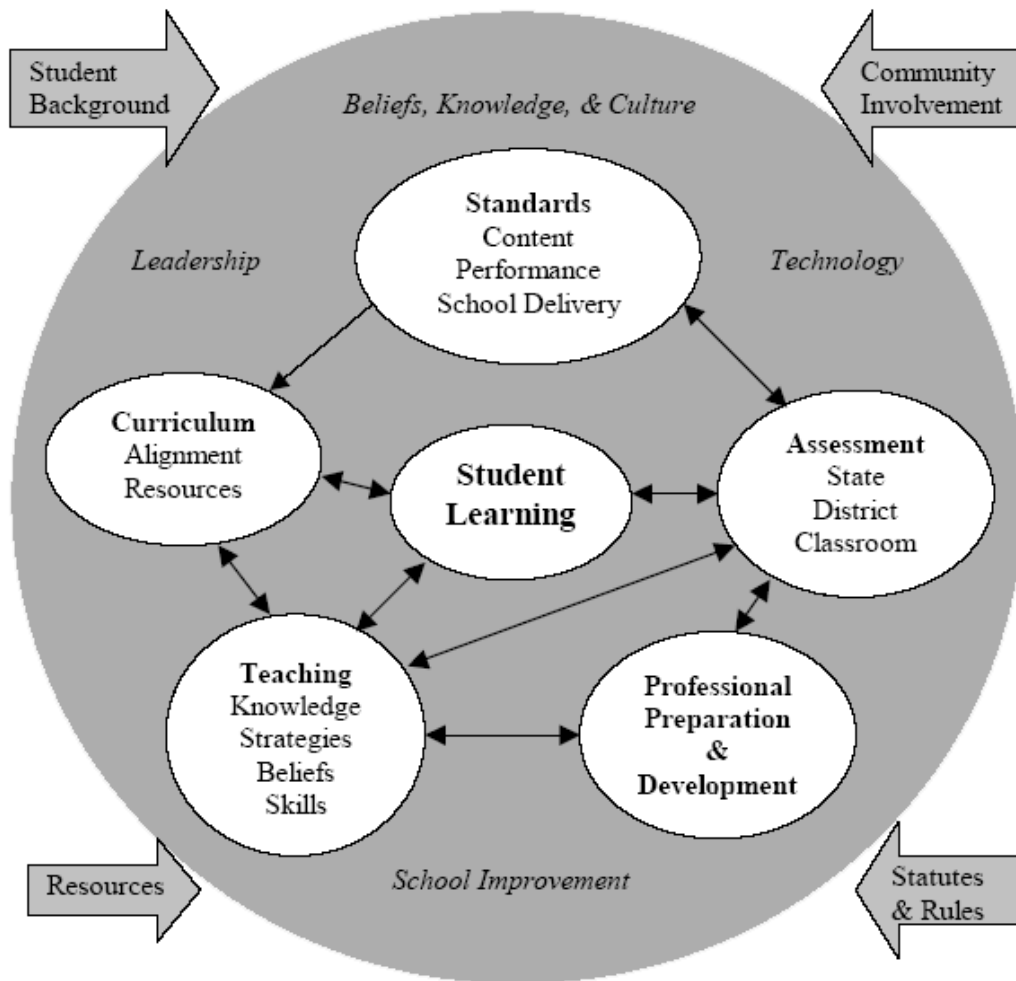
Figure 1. The Standards-Based System in Wyoming: Ensuring Equality of Educational Opportunity.

As depicted in the model, student learning is the focus of the entire system. All components must relate to improving student achievement. Standards and assessments are crucial components of this approach, and while well-designed standards and assessments are necessary for this approach to be successful, they are far from sufficient. Teaching and leadership, including pre-service and in-service education, are vital for improving student achievement. Having all of the components of the model commonly focused or aligned can

positively influence student achievement. On the other, having some of these components work at cross-purposes will likely not produce the desired outcomes.

In a standards-based system, assessments are closely linked to standards, and assessment results are reported in terms of the performance standards. The assessments provide information about the achievement level of the students and help determine which performance level they have reached. Most importantly, they inform instruction. Assessment should involve a range of strategies appropriate for drawing inferences about individual students, classrooms, schools, and districts.

Armed with data on how students perform according to standards, schools and districts must make the instructional changes needed to improve performance. It is critical that schools and districts monitor the curriculum and instructional practices of teachers to determine if students are provided the "opportunity to learn" the standards they are expected to meet. Districts and schools should use this information to require and support improvement of instruction and learning in every classroom.

Professional development should focus on the standards for student performance. Districts should use the results from student assessments and information about curriculum and instructional practices to design their professional development programs. Research on professional development indicates that focusing on the content and curriculum teachers teach, rather than special topics like cooperative learning, are the most effective (Cohen & Hill, 1998). Activities centered on student assessment also appear to have an impact on changing instructional practice. Content-, curriculum- and assessment-based professional development also appears to affect student learning: one research study indicates that students whose teachers participated in these types of topics outperformed others on state assessments (Cohen & Hill, 1998).

When standards-based educational reform was first articulated (c.f., Resnick & Resnick, 1992), school delivery standards were an important part of the mix. However, as many state and national reform policies began to take hold in the 1990s, school delivery standards appear to have fallen by the wayside. These standards place a great deal of responsibility on those outside of the classroom such as district and state officials and policymakers to ensure that teachers and students have coherent policies, support, and resources necessary for high quality teaching and

learning to occur. It has been easier for politicians to place the responsibility for change on the shoulders of those inside the classroom than to share in the responsibility for change. For example, in spite of the many policies around the country to provide rewards and sanctions for students and teachers, there are no policies that we know of where legislators and other state or national policymakers suffer consequences if student achievement does not improve. Yet, the policies and support (or lack thereof) have a direct bearing on what happens in the classroom. In a true standards-based model, all members of the educational community bear some responsibility for improving student learning. It is unfair to place all of the responsibility on those inside the classroom because they often have the least control over the structure of their endeavors.

What is an Assessment System?

Obviously, an assessment system is comprised of individual assessments. "But a collection of assessments does not entail a system any more than a pile of bricks entails a house" (Coladarci et al., 2000). An assessment system is a well-articulated set of assessments, each of which contributes toward supporting inferences related to the identified purposes of the system. A well-designed system will support inferences where the whole (the system) is greater than the sum of the individual assessments. An assessment system, therefore, has the following key characteristics (Coladarci et al., 2000):

- ➢ **The set of assessments comprehensively addresses the content and performance standards or other well-defined learning targets.** An assessment system provides evidence about student achievement directly related to the Wyoming Content and Performance Standards. Because the content standards are broad, the assessment system should be able to provide information at a more detailed level such as benchmarks and course outcomes.

- ➢ **The assessment system provides students with multiple opportunities, using multiple formats, to demonstrate their knowledge and skills related to the Wyoming Content and Performance Standards**. In other words, the system should allow *students the opportunity to show what they know*, but the system should be designed in such a way so

that students who have not mastered the standards should not be able to pretend to know (e.g., not "pass" the assessments).

➢ **Each of the individual assessments has a clear and explicit rationale.** When an assessment is included in the system, its purposes should be made clear, and its connection to the rest of the system should be part of a logical and coherent plan.

The most important characteristic distinguishing an assessment system from a simple collection of tests is that a system is designed to provide a cohesive array of information on student performance. The various components of the system across the different educational levels provide complimentary information so that decisions can be based on valid inferences.

A comprehensive assessment system often includes multiple components targeting local, state, and national levels of the educational system (see Figure 2). The overlap among levels indicates that assessments at one level may also be used at another level. For example, a good individual classroom assessment may be "scaled-up" for use at the school or district level, as well. Also, notice that classroom assessments make up the largest portion of the assessment pyramid.
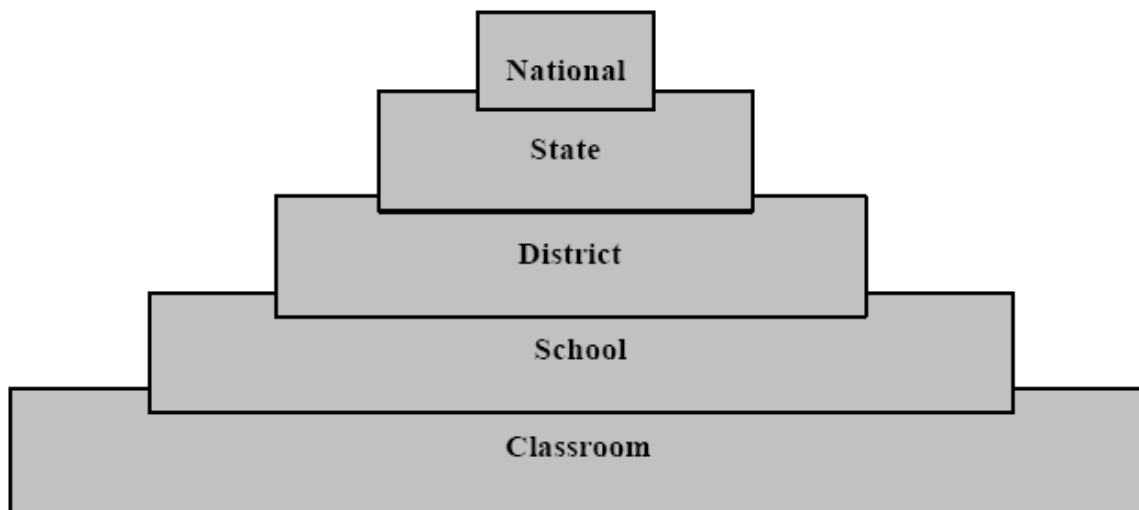
Figure 2. Levels of Assessment.

Assessments can be broadly classified as internal or external to the classroom. Assessments internal to the classroom are the most influential for making a difference in student learning while external assessments are often used for program evaluation or for accountability. Teachers need routine assessments and quick turnaround of data in order to make sound instructional decisions for their students. Classroom teachers must regularly use classroom assessments to measure on-going progress in the mastery of standards.

When the "close-up" data derived from classroom assessments are used in conjunction with the large-scale snapshots of state and district level exams, an accurate and comprehensive picture of student and school achievement emerges. For this reason, any assessment system designed to support student learning should include multiple components to produce a valid system. It is crucial that the assessments across the various levels, i.e., state, district, and classroom, work cohesively together to send a similar message about expectations for teaching and learning. If all levels of assessment are based on the same content standards, the "message" is more focused and all levels of testing can influence classroom practice.

Assessments at each of the levels contain various strengths and weakness, depending on the purpose(s) of the assessment. These advantages and disadvantages should be considered when creating a district assessment system so that assessments are used for the purposes for which they offer the most positive and fewest negative effects. Purposefully selecting assessments from different levels can lead to an assessment system that supports student learning while meeting important accountability requirements.

CHAPTER II: ASSESSMENT SYSTEM DESIGN PRINCIPLES

The purpose of this chapter is to provide a discussion of the conceptual underpinnings of the principles most relevant for designing a district assessment system. Although this chapter is written as non-technically as possible, the discussion presented here is targeted to those readers with at least some background understanding of basic assessment principles. For those wishing for a more thorough treatment of some of these topics, we have tried to suggest appropriate references. A working understanding of these principles is crucial in order to develop valid district assessment systems, and we urge the reader to take the time necessary to become familiar with these concepts presented in this chapter.

Overarching Design Considerations

There are three considerations--purpose, use, and validity--that subsume all of the specific principles for designing any assessment system. However, before discussing these overarching concepts, we would like to discuss one aspect of standards-based education that is often misconceived as standards-based programs are implemented.

Standards are big ideas! When the American Association for the Advancement of Science (AAAS) first published *Science for All Americans* (Rutherford & Ahlgren, 1989), its intent was clear—the study of science had become fragmented and students never really gained a sense of the organizing concepts in science. *Science for All Americans* served as the foundation for the National Science Education Standards (NRC, 1996), which organized the study of science into six broad organizing themes (standards). Relying on the structure of the discipline (science in this example) to organize instruction is supported by the latest theories in human learning. Disciplinary content and our schema for organizing knowledge are usually hierarchical with major concepts and principles subsuming more specific facts. Learners who construct accurate schema when learning the concepts and processes of a discipline can better store those concepts in long-term memory and use those concepts and processes in novel situations (Anderson, Reynolds, Schallert, & Goetz, 1977). Subject matter is taught, too often, as a series of isolated

facts, and students are unable to develop an accurate schema, and they cannot develop a sense of the discipline (Rutherford & Ahlgren, 1989).

The National Council of Teachers of Mathematics (NCTM, 1989) made a similar argument when they published the first set of mathematics standards. While teachers need to break apart these big ideas in order to teach some of the foundational concepts, they and their students should learn to have a sense of the organizing framework of the discipline so the concepts and skills can be more readily called upon for later use. While this handbook is not designed to provide details about implementing standards-based instructional practices, the assessment practices we use have a tremendous influence over how content and processes are taught.

If assessments are focused on narrow and isolated benchmarks, and if students are expected to "master" these narrow skills prior to receiving subsequent instruction, classroom instruction will likely be focused on isolated concepts and skills as well. This notion of teaching small amounts of information and testing at each step is classic behaviorism (Shepard, 2000), and this theoretical approach to student learning is no longer considered an effective method for promoting deep and meaningful learning (Shepard, 1991). However, because it is easier to construct assessments of small, specific concepts, many assessments are built in ways that inadvertently support behaviorist approaches of teaching (Shepard, 1991). Designing assessments that measure integrated concepts and skills are more challenging but can better promote student learning of challenging content. **Therefore, an overarching principle for the design of assessment systems is that the assessments comprising the system are focused on "big ideas" and not on isolated sets of skills.** This does not mean that teachers should avoid breaking things into component pieces for students, but when doing so, they should keep the big ideas in mind. Teachers, by regularly returning to the big picture, will help students generate these big ideas. Without this guidance and by focusing on the smaller pieces of knowledge, it is unlikely that students will be able to generate accurate schema on their own.

Purposes

The first step in designing any assessment or assessment system is to specify the purposes. Many times purposes are assumed or implicit, but that often causes trouble down the line when one learns that two or more groups held contradictory assumptions about the reasons for conducting the assessment. Purposes must be made explicit to all of the key stakeholder groups, and these stakeholders should help frame the purposes. The purposes, made explicit and prioritized, will serve as touchstones when facing difficult design decisions down the road. Certain design choices will serve some purposes better than others and being able to return to the purposes will help bring clarity to these design decisions.

The legislature and the State Board of Education have given us some purposes already. Specifically, there is a set of purposes that fall under the heading of accountability such as certifying that students have the knowledge and skills necessary to graduate from high school, that the high school diploma "means something," and that school districts are providing an opportunity for all Wyoming students to meet the same high standards.

There is another set of purposes related to improved teaching and learning in Wyoming's classrooms. These would include such things as having information necessary to modify instructional decisions for students or receiving information that will allow school leaders the opportunity to evaluate the effectiveness of various programs. Having an assessment system structured so that the assessment results can be used to identify students in need of remediation and to identify professional development needs are other viable purposes.

Districts can choose to design a system to fulfill the accountability purposes only, but that would be a very inefficient use of time and resources. Therefore, some other purposes mentioned above, especially the improvement of teaching and learning, should be part of a district's system. Districts will have to prioritize the stated purposes in order to conceptualize the overall design. For example, if accountability were the highest priority purpose, then a system based on tests at key checkpoints throughout students' K-12 career would suffice. However, this type of system would probably not do as much for providing the information necessary for improving teaching and learning as a system that incorporated more classroom components. In other words, the specific purposes and uses drive the design of the system, so it is worth taking the time necessary

to build consensus among key stakeholders in order to have a system that can best meet the stated purposes.

The overarching district system fulfills multiple purposes, relying on different types and levels of assessments to meet different needs. For example, including national and state tests as part of the system can meet accountability purposes, but classroom assessments are crucial for the improvement of teaching and learning.

Using Assessment Results

Assessment results provide important information to all facets of the school community. Policymakers, administrators, teachers, students, and parents all use assessment information for a variety of purposes. Collectively, these users make decisions about whether schools are functioning effectively for an individual child or for all children as a whole. A district assessment plan needs to outline how the results of each type of assessment are used because validity can only be evaluated in terms of the use(s) of a particular assessment.

At the district level, assessment results should help make informed decisions about curriculum and instruction. Since these are supposed to be aligned to content and performance standards, assessments can help indicate a need for revising curricula or instructional methodology. The PAWS and district assessment results can help target specific areas necessary for district improvement.

At the school level, assessment results should be used to help determine school improvement goals. They may also indicate educational strategies for improvement in targeted areas. Assessment results are one of several items that should be used for the identification of budget priorities. At both the district and school levels, results should be used to monitor the effectiveness of staff development activities and at-risk programs.

The following chart outlines additional uses of assessment results at three different levels; the instructional level, the instructional leadership/support level, and the policy level. When considering how the results of each type of assessment within the district assessment plan will be used, the uses described mainly in the instructional leadership/support level and the policy level should be considered.

INSTRUCTIONAL LEVEL	
Users	Uses
Student	• Track own Success: *Am I meeting the teacher's standards?* • Identify own needs: *What help do I need to succeed?* • Connect effort to results: *Does my work pay off?* • Plan for educational and vocational needs: *What will be the next steps in my learning?*
Teacher	• Identify needs of individuals: *What does this student need help with?* • Identify needs of class or group: *What do these students need help with?* • Group students: *Who among my students should work together?* • Grade: *What grade should appear in the report card?* • Evaluate Instruction: *Did my teaching strategies work?* • Evaluate self: *How do I need to develop to be a better teacher?*
Parent	• Track child's success: Is *my child succeeding in school?* • Identify needs: *What does my child need in order to succeed?* • Evaluate teacher: *Is my child's teacher doing the job?* • Evaluate school: *Is this school (district) working for my family?*
INSTRUCTIONAL LEADERSHIP / SUPPORT	
Principal	• Evaluate instructional program*: Is instruction in particular areas producing results?* • Evaluate teachers: *Is the teacher producing results? Does the teacher meet minimal performance standards?* • Allocate resources: *How shall we spend building resources in support of instruction?* • Determine professional development needs: *What kinds of professional development would help this/these teacher(s)?*
Lead Teacher/Dept. Chair	• Assist new teachers: *What does this teacher need to assure competence?* • Support instructional program: *Which teacher(s) need what help to do the job?*
Counselor/ Psychologist	• Identify students with special needs: *Who needs (can have access to) special support services such as remedial programs?* • Match students to program: *What student should be assigned to which teachers to optimize results?*
Curriculum Director	• Evaluate program quality: *Is the program in a particular area of instruction working?*

Figure 3. Users and Uses of an Assessment System (From Stiggins, 1997).

POLICY LEVEL	
Central Admin	• Evaluate program: *Are programs producing student learning?* • Evaluate principals: *Is the building principal producing results?* • Allocate resources: *Which schools need/deserve more or fewer resources?*
School Board	• Evaluate program: *Are students in the district learning?* • Evaluate superintendent: *Is the superintendent producing results?*
State Dept. of Education	• Evaluate program: *Are programs across the state producing results?*
Citizen/ Legislator	• Evaluate program: *Are students in our schools achieving in ways that will allow them to be effective citizens*?

Figure 3, cont'd. Users and Uses of an Assessment System (From Stiggins, 1997).

Validity

Validity is the most important technical consideration in assessment design and implementation. In fact, all of the other design principles, if followed, contribute toward building a valid assessment system. Validity is often defined as the degree to which a test measures what it is intended to measure. But it is not the test that is valid or not; it is the inferences—in the context of a particular use—that are valid or not. This might seems like a trivial distinction, but it has important implications for evaluating the validity of an assessment system.

An evaluation of a district's assessment system would, in essence, be a validity investigation. A validity study would NOT result in a single "validity coefficient." Validity coefficients are often part of predictive validity studies and are not sufficient to support the validity of the inferences from the assessment system. Validity should be thought of as an argument where logical, theoretical, and empirical information is marshaled to present the best case (Appendix K contains several possible lines of inquiry for validity studies).

Carrying out a complete and detailed validity investigation is probably prohibitive for most Wyoming school districts. However, validity concerns should be used to frame all of the other required inquiry (e.g., alignment, consistency) so that these data can also be used to build a validity argument.

Primary Design Principles

The overarching goal of the district assessment system is that it should be able to support inferences about students' mastery of the common core content areas and the districts' capacity for ensuring equality of educational opportunity. In order to meet this laudable goal, specific design principles are necessary. The design principles should guide the development of districts' Body of Evidence systems in order to make sure the system could possibly fulfill the stated purposes. Focusing on the highest priority design principles first—alignment, consistency, standard-setting, and fairness—should lead to the development of solid district assessment systems and should enable district designers to concentrate their energy on the most important concepts. In the following section, we present the conceptual underpinnings of each design principle.

<u>Alignment</u>

Alignment is an often-misunderstood term in educational measurement. In the broadest sense, it is the core idea that in a standards-based system, all parts of the system (i.e., assessments, teaching strategies, instructional materials, professional development) work together to ensure student learning of the content standards. In terms of the relationship between standards and assessment, alignment refers to the match between the items on the assessments and the knowledge and skills represented by the curriculum and the standards. While this sounds fairly straightforward, alignment is often addressed incorrectly because alignment is much more than simply matching the content of the items with specific standards. There are many layers to the concept of alignment.

The alignment of the content of a district's assessment system to the standards is at least a **two-way process**. It is not sufficient to assume alignment because all items on assessments can be matched to one or more standards. Districts must also assure that all the standards are adequately assessed. In the Venn diagram below (Figure 4), many of the items on all the

assessments match the standards. However, fewer than half of the standards are assessed. In a well-aligned system, the two circles below would overlap almost perfectly.

Another way of visually representing alignment is presented in Figure 5. The major elements of the content contained within a standard should be considered and represented with the items selected.
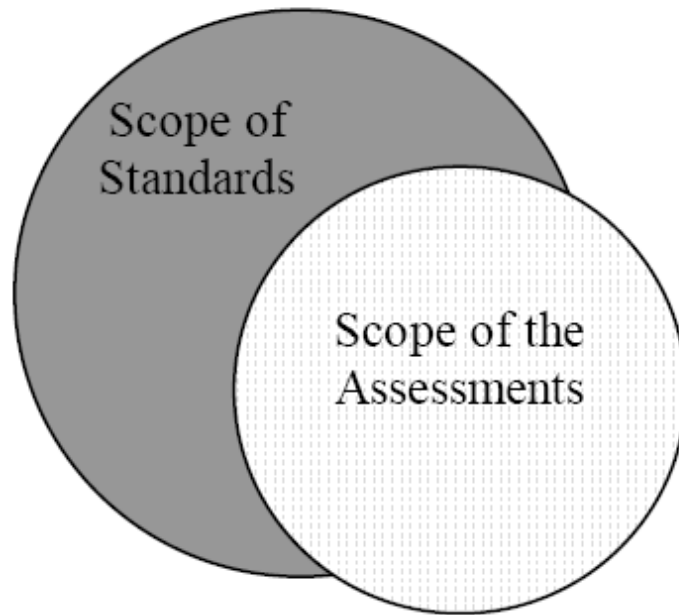
Figure 4. Overlap of Standards and Assessments (from U.S.D.E, 2000).
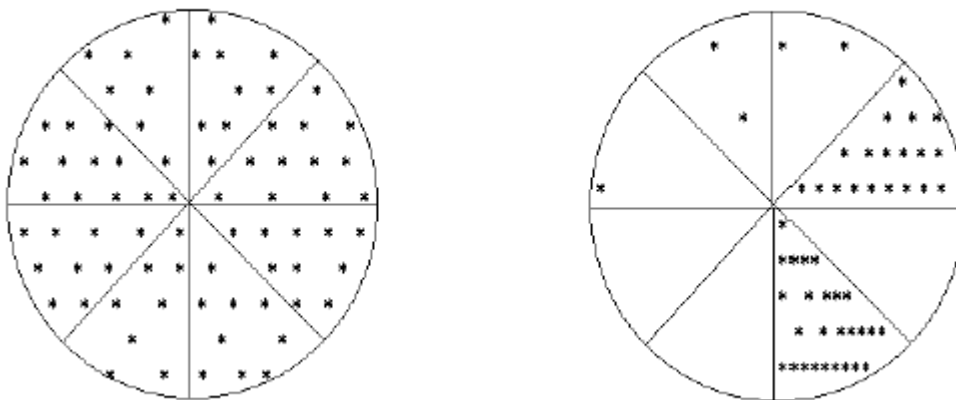
Figure 5. A Well-Sampled and Poorly-Sampled Standard.

In the circle on the left, the standard, represented by the circle, is well-sampled. The sections represent the major elements of the standard. The test items, represented by the *s, are distributed fairly evenly throughout all of the major elements. On the other hand, the circle on the right illustrates a standard that is not well-sampled. The majority of items are within only two of the eight major elements of the standard. Some of the major elements have no items at all. Inferences drawn from the resulting data will likely under-represent the construct and not be very valid. Webb's (1999) framework for analyzing alignment is useful for explaining the complexity of alignment.

Categorical concurrence. *"The categorical concurrence between the standards and assessment is met if the same or consistent categories of content appear in both documents"* (Webb, 1999, p. 7). If the assessments that comprise the assessment system are faithfully built directly from the standards, the categorical concurrence between the assessment and the standards can be assured.

Balance of representation. "*The balance-of-representation criterion is used to indicate the extent to which items are appropriately distributed across standards"* (Webb, 1999, p. 9*). All standards are not created equal, but it is up to the district design team to appropriately weight the standards by assigning more items/tasks to the higher priority standards. Nevertheless, this criterion calls for the assessment items/tasks to match the standards in the proportions expected.

Depth-of-knowledge consistency. *"Depth-of-knowledge consistency between standards and assessments indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards"* (Webb, 1999, pp.7-8)." This criterion calls for the alignment to be judged not only in terms of the category matches between the test items and the standards, but also in terms of the performance levels established for the assessments. This is generally the most overlooked alignment component. It requires careful analysis to determine if the assessment is accurately targeting the cognitive levels called for in the standards.

Range-of-knowledge correspondence. *"The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the*

assessment items/activities. The criterion for correspondence between span of knowledge for a standard and the assessment considers the number of objectives within the standard measured with at least one related assessment item/activity" (Webb, 1999, p. 8). This criterion indicates that alignment needs to occur at a finer grain level than simply matching tests items to broad standards. The alignment should occur at the benchmark level or finer grain to ensure that the standards are appropriately sampled. This does not mean that each benchmark should be assessed separately, rather it simply means that the assessments should be built from a blueprint that reflects the appropriate weight of each benchmark, but the decision can certainly be made at the standard level.

Categorical concurrence and balance of representation can be established by using two matrices or maps between the standards/benchmarks and the various assessments. The first matrix would focus on the match between standards and assessments (see Chapter III, Figures 9 & 10). Then a finer grain matrix should be developed for each major assessment that documents the match between assessment items/tasks and the various benchmarks assessed (see Figure 6 below). These matrices also ensure that alignment is established through a two-way process where the items are matched against the standards and the standards are checked to make sure they appropriately represent the items.[1]

Using the finer-grained matrix pictured below (Figure 6), the assumed cognitive demand can be included on this blueprint to make sure the design team is at least trying to target higher-order thinking skills. Once these items/tasks are created, a careful, cognitive analysis of the test items and the performance standards should be conducted. Finally, perhaps the best source of evidence to ensure depth-of-knowledge consistency can be collected using "think-aloud" protocols to truly check the way that students are interpreting the cognitive demand of the items. This last source of data is beyond what many districts will be able to collect, but think-aloud protocols are worth pursuing because they provide tremendous insight into the quality and the meaning of the assessments.

---

[1] See Coladarci, et al. (http://mainegov-images.informe.org/education/g2000/measured.pdf) for some practical approaches for evaluating alignment.

Standard	Benchmarks	Number of Items/Tasks or Score Points			Totals
		Cognitive Demand			
		Low	Moderate	Complex	
Reading	1	1	3	2	6
	2	1	1	1	3
	3	2	2	2	6
	4	2	1	1	4
	5	3	4	5	12
	6	1	2	1	4
	7	1	4	6	11
	Totals	11	17	18	46

Figure 6. An Example of a Test Blueprint for a Language Arts (Reading) Assessment. The number of items/tasks/score-points in each cell of the matrix is just an example and not a specific recommendation.

Consistency

Reliability and consistency are often used interchangeably, but we use consistency here to indicate the focus is on the assessment system. In other words, the set of assessments should yield consistent decisions about students' performance related to the standards. Like validity, it is the collection of assessments that needs to yield decision consistency. **In terms of the Body of Evidence graduation requirement, the importance of having a system that yields decision consistency cannot be understated because one of the intentions behind using the Body of Evidence approach was to reduce the number of misclassified students compared to using a single test.** On the other hand, it is often tempting to focus evaluations of assessments and assessment systems on reliability because it can be quantified easier than many of the other design principles. We should resist concluding that simply because we can compute coefficients, reliability is more important than other considerations.

There are several sources of error in an assessment system that can affect the consistency of the decision. Many of these sources of error are related to individual assessments, but there are also errors that are associated with the system itself. For example, changing the weighting

scheme for combining results from the various measures in the system would certainly affect the consistency of the decisions (this would obviously affect the validity as well) and needs to be investigated. Additionally, the consistency of the decisions can be evaluated by having a different panel rate the same evidence and by evaluating the effect of slightly different decision rules on the consistency of the decision.

Beyond the panel and judgment issues, there are not well-specified methods for evaluating the consistency of an assessment system. Therefore, most of the consistency evaluations focus on individual assessments. Many introductory measurement textbooks can provide guidance for choosing the most appropriate means for conducting these analyses, but district personnel are urged to at least speculate, based on data from individual assessments, about the reliability of the system.[2] Reliability is positively influenced by increasing the number of test items (all other things being equal). If a system includes many assessments, then, in fact, the system will be comprised of at least several hundred items or tasks. Therefore, although not guaranteed, if assessments are well aligned with standards/benchmarks and enough measures are used to make a decision about a student's competency in a given content area, then consistency should not be a problem. However, this assurance is dependent on the level of the decision. Using the Body of Evidence system as an example (see Chapter IV), if the pass/no pass decision is made at the content area level (using a compensatory approach), this assurance is accurate. On the other hand, if the pass/no pass decision is made for each benchmark, then the reliability of the system is only as good as the reliability of the least reliable benchmark assessment.[3]

---

[2] The following sources will provide guidance for evaluating reliability of individual assessments: Airasian, P. W. (1994). *Classroom assessment (2nd ed.).* New York: McGraw Hill. Linn, R. L. & Gronlund, N. E. (2000). *Measurement and evaluation in teaching (8th ed.).* New York: Macmillan. Popham, W. J. (1999). *Classroom assessment: What teachers need to know (2nd ed.).* Boston: Allyn & Bacon. Salvia, J. & Ysseldyke, J. E. (1998). Assessment (7th ed.). Boston: Houghton Mifflin.

[3] One way to speculate about the reliability of the system based on the reliability of the individual assessments is to take the item-level results from several components of the BOE and put them together in a statistical package such as SPSS or Excel and then treat these as a "single exam." Then using one of the reliability procedures (e.g., KR-21 or Cronbach's alpha) described in one of the texts named in the previous footnote, the reliability of the system can be estimated. Again, referring to Coladarci et al. http://mainegov-images.informe.org/education/g2000/measured.pdf will be helpful.

Fairness

The assessment system and the individual assessments could be designed inadvertently to be unfair to certain subgroups of the population. There are two parts to this issue. The first, an assessment issue, focuses on whether the set of assessments leads to unfair inferences about students' performance on the standards. The second, a policy, legal, and moral issue, is concerned with the implications based on the analyses of the assessment issues.

Obviously, an assessment system should not be biased toward any groups of students. Ideally, it should not be biased against or for any individual student but that is quite difficult to evaluate. Trying to ensure that a fair system is being used should be addressed in the design phase of the assessment system. The following questions are examples of those that should be addressed while designing the district system. You may notice that, for good reason, many of these questions are validity questions. Fairness is certainly subsumed by construct validity (Linn, Baker, & Dunbar, 1991; Messick, 1989) and findings of bias indicate that the test(s) is differentially valid for the subgroups of students in question.

➢ Are appropriate accommodations being used so that educators are able to make valid inferences about the achievement of students with disabilities?

➢ Do the assessments contain excessive language demands so that second language learners or other students with language difficulties are unfairly denied the opportunity to show what they know?

➢ Is a variety of assessment formats employed so that students who tend to perform better or worse on specific formats are not treated unfairly by the system?

➢ Do the prompts, stories, and examples used in the assessments contain text that unfairly favors one group of students over another? For example, if all of the extended-response problems in mathematics were based on examples from sports, some could justifiably argue that the assessments are biased in favor of boys over girls.

These are just a few examples, but the point is that it is easier to address these during the design process than after the fact. Even if these issues have been addressed up-front, it does not relieve us of the responsibility of checking for bias after the data have been collected.

The results of all assessments used in the assessment system should be disaggregated for identified subgroups of students. Examining patterns in the scores on the different assessments in the system might allow you to flag particularly problematic assessments.[4] Disaggregated results can also allow for the analysis of achievement for different subgroups on the assessment system. However, simply uncovering differences in score patterns among two or more subgroups does not indicate that the assessment system is biased. Further, the assessment system should not be designed to eliminate or mask these group differences. This leads to the second issue introduced above.

Group differences on assessments are more often indicative of differences in opportunity-to-learn (OTL) than of a biased assessment system. It is often tempting to "shoot the messenger," but differences in scores on an achievement test often point to a much deeper and pervasive problem. If group differences in score patterns are evident, the first set of questions one asks should focus on checking the assessments for problems of bias. Once we are satisfied that the assessment system is not biased, there are a different set of questions related to OTL that needs to be addressed. For example:

> Is one group of students systematically denied the opportunity (either through lack of encouragement or other means) to participate in higher-level classes?

> Is the curriculum designed from a single cultural or socioeconomic point of view so that certain students are not provided the same learning opportunities as others?

The disaggregated assessment results should be used to improve educational programs and ensure that all students have a chance to learn the required knowledge and skills as well as succeed on the assessments.

Finally, if group differences are minimal, but many students are still not meeting the standards, it could be that most students in the district are not being prepared to meet the

---

[4] The first post-hoc step measurement professionals take to search for item bias is to examine items/tasks for differential item functioning (DIF). We do not suggest that districts conduct this type of statistical analysis—it is too time-consuming and complex.

standards and therefore are not being provided a fair opportunity to learn the standards. This kind of information can be used to realign curriculum and instruction as well as examine other issues related to student performance to ensure that students are really being provided the opportunity to learn Wyoming's Content and Performance Standards.

Standard Setting

Standard setting is the process of determining the scores that divide various performance levels (i.e., cutscores) on an assessment. A challenging aspect of standard setting is that many methods exist to set standards and establish cutpoints. Nevertheless, all of the "best-practices" methods rely on some type of deliberative process for arriving at cutscores rather than simply establishing performance levels through the use of traditional percentages or other means. It is beyond the scope of this handbook to provide a comprehensive description of standard setting methods, rather our intent is to simply introduce the topic.[5]

The most important characteristic of all standard setting methods considered for use on a Body of Evidence system is that cutscores should be tied closely to proficiency descriptions, and these descriptions should be widely shared and agreed upon. Further, the evidence of any process used should be documented. **In other words, there should be an agreed-upon description**

---

[5] The Wyoming Department of Education, as well as, the Wyoming Activities Assessment Consortium have offered standard setting workshops over the span of four years in 2000, 2002, 2002 and 2003. For information on standard setting, we suggest the following: Cizek, G. (2001) (Ed*.). Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Publishers. Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000a). *Handbook for setting standards on performance assessments*. Washington, DC: Council of Chief State School Officers. Jaeger, R. M. (1995). Setting standards for complex performances: An iterative judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, 14(*4), 16-20. Kahl, S.R., Crockett, T.J., DePascale, C.A., & Rindfleisch, S.L. (1995, June). Setting standards for performance levels using the student-based constructed-response method. Paper presented at the annual meeting of the American Research Association, San Francisco, CA. Kingston, N. M., Kahl, S.R., Sweeney, K., & Bay, L. (2000). *Using the Body of Work Method to Set Performance Standards: Experiences in Multiple States*. Measured Progress Research Report 2000-1. Dover, NH. Plake, B. S. & Hambleton, R. K. (1998). A standard setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work. Paper presented at the annual meeting of the American Research Association, San Diego, CA. Shepard, L.A.(1994). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In Procedings of *Joint Conference on Standard Setting for Large-Scale Assessments*. Washington, DC: National Assessment Governing Board and
National Center for Educational Statistics.

stating what it means to pass or be proficient (or whatever level) and there should be evidence that the cutscore reflects this description.

Standard setting methods can generally be divided into item-based methods and student work-based methods. Because of the nature of the Body of Evidence approach, we believe work-based methods are more appropriate. The Body of Work method (Kahl, et al., 1995; Sweeney, Kahl, Kingston, & Bay, 2000) is one such method that we think could be adapted quite well to set standards on a Body of Evidence system or other aspects of a district assessment system. The example presented in Appendix A describes how a district could use the Body of Work method to establish cutscores on a Body of Evidence assessment system.

Comparability

The Wyoming Supreme Court's 1995 decision of school funding was largely focused on ensuring that all students across Wyoming receive equal educational opportunities. The new graduation requirements mandated by W.S. 21-2-304 were written to ensure that all Wyoming students reached a certain threshold of competency prior to graduation. Therefore, the notion of comparability is important in designing a district assessment system. **We recognize that comparability cannot be assured statewide without using a common set of assessments, but by following these design principles, district personnel should be able to detect whether or not students have reached the threshold implied by in the Supreme Court's decision.** Comparability within districts, across classrooms and schools, both within the same year and across multiple years needs to be ensured.

Within a given year, it is crucial that a student's chance of meeting the standards is not dependent on which school within a district a student attends or which set of teachers within a given school a student draws. One way to ensure comparability across teachers and schools is to have at least some common assessments (broadly speaking to include such things as projects and activities) that comprise the assessment system.

Comparability across years is a more difficult issue. Simply using the same assessments from year to year does not ensure year-to-year comparability. There will likely be some memory of assessments from the previous years, therefore using the same assessments year in and year out will actually make these assessments easier. Therefore, items/tasks should be replaced on

individual assessments after they have been used no more than a few times. These new items/tasks should be developed from the same matrices discussed previously to help maintain comparability. Unfortunately, the most appropriate methods for ensuring year-to-year comparability require fairly sophisticated equating designs. This is not an expectation for Wyoming school districts, but districts should have a plan for maintaining year-to-year comparability.

Multiple Measures

'Multiple measures' is an often used but commonly misconceived term (Olson, 2001). While some states appear to believe that multiple measures can be defined by allowing students to take the same test multiple times, we, in Wyoming, do not believe that is the case. Multiple measures should be used to increase the validity, alignment, consistency, and fairness of the assessment system. This section highlights some of these ideas and clarifies the rationale behind using multiple measures.

The matrix of multiple measures should be used to enhance the technical qualities of the assessment system, not simply to add additional assessments. If the additional assessments do not contribute to improving the alignment, validity, or fairness of the system, they should not be used. Adding additional measures will almost always improve the reliability, so that should not be used as the criterion for determining whether or not to employ an additional measure. The multiple measures should, whenever possible, represent different formats and strategies to enhance the fairness and alignment of the system.

Determining the best approach for combining multiple measures to yield the most valid inferences is a complex technical issue. In fact, it is one of the most troublesome issues facing national measurement experts. In spite of the lack of "answers" from the experts, there are a few key guidelines. The combination of multiple measures should reflect a justifiable and logical approach rather than a simple mechanical approach (i.e., a simple average). A simple mechanical approach might be the best approach, but it should be determined rationally and not by default. Additionally, the combination of multiple measures should reflect the importance of each measure in terms of its coverage of the standards and the quality of the measures.

How many measures are enough? Unfortunately, there is no simple answer. Anything more than one is considered multiple. Simply stated, enough measures should be employed so that the alignment and fairness criteria are improved over a single assessment. In general, addressing the alignment and fairness principles will lead to an appropriate use of multiple measures and combining multiple measures is essentially a standard setting exercise.

Credibility

Many would argue that the public support for higher standards is due to perceptions that many students are leaving high school without the knowledge and skills necessary for success. Whether this is true is not the question; rather the issue is how to build credibility in our educational system. The assessment system and the resulting decisions (inferences) should lead to an increase in confidence in the school system by the various stakeholder groups, such as key business people, local legislative representatives, parents, and other key community members.

There are many ways to build confidence in the assessment system. First and foremost, members of key stakeholder groups (e.g., parents, certain businesspeople) should be invited to participate in discussions about the system and even to be included as members of a design team or a review panel. Another way to build credibility is to share student work with members of the public. This can be accomplished through public exhibitions or through publishing tasks and samples of student work. If the public can see the type of work expected of students, it will likely build credibility in the system.

Consequences

The social consequences of the assessment system are certainly subsumed under validity, but bear emphasizing separately here. There are consequences when any new policy or practice is enacted. Obviously, we always intend for these consequences to be positive. Unfortunately, many unintended, negative consequences result from policy decisions. **Therefore, each district should collect and analyze data to evaluate the consequences of the system, both intended and unintended. As part of this investigation, the district should evaluate whether the stated purposes are being served by the system and whether the system needs to be**

modified to best serve the stated purposes. The district should also take appropriate action to improve the system and minimize the negative, unintended consequences.

Consequences cannot be evaluated until after at least a couple of years of data have been collected; however, districts should build a data collection plan up front that will allow them to collect the data necessary for evaluating consequences. For example, if a district were evaluating the effect of the Body of Evidence system on course-taking patterns, then data should be collected prior to fully implementing the Body of Evidence system and followed up in subsequent years. Districts should also collect data to document that students are not being systematically denied opportunities to learn the standards and meet graduation requirements.

Summary of Design Principles

The concepts presented in this chapter provide a foundation for designing assessment systems. They can be applied to the design of the overall district assessment system and to the design of specific aspects of the system such as the Body of Evidence system. While this chapter might appear overwhelming, it should be used as a reference when working through the more practical chapters that follow. As mentioned previously, the design principles are not all created equal, and they need to be prioritized as the system is being designed. In general, alignment should be the starting point after the uses and purposes have been clarified. Then other principles such as consistency and fairness should be incorporated as each assessment is designed and built. The following chapters provide specific and pragmatic details for designing assessment systems.

CHAPTER III: DESIGNING A DISTRICT ASSESSMENT SYSTEM

District Assessment Systems' Place in the Reform Model

The major purpose of the district assessment system is to document that all students are being provided a fair opportunity to learn the Wyoming Content and Performance Standards. Equality of educational opportunity, as established by the Wyoming Supreme Court in 1995 is not based solely on simply equalizing inputs across the system; rather it must also be judged by evaluating educational outcomes in a comparable fashion. A valid assessment system can maximize learning by informing educators, policy makers, students, and the public about student achievement of the standards, thus helping all stakeholders from the student to the legislator make sound decisions about education and learning.

Many programs, such as accreditation and Title I, have specific assessment requirements for districts and indicate how assessment results are to be used. The district assessment plan should be the umbrella plan for all assessments administered and/or collected at the district level. A well-planned, comprehensive district assessment system can be constructed to fulfill all of the various district assessment requirements using an efficient number of assessments. The purpose of this chapter is to provide guidance for constructing the overarching district assessment system to meet the Wyoming district accreditation requirements and to provide a framework for incorporating the Title I and Body of Evidence assessment systems within this overarching system.

Creating a District Assessment Plan

Some Initial Considerations

Sampling. All of the design principles presented in Chapter II are applicable for designing the district assessment system. However, the concept of sampling deserves special attention for the design of district systems. Assessments should be purposeful. Logical rationale should be applied when selecting grade levels and content areas to be assessed and the types of

assessments to be used. The concept of sampling can help reduce the number and amount of time devoted to district-level assessments, yet still yield adequate data for school improvement and accountability. Classroom assessment should be continuous and used to inform instruction on a very regular basis, but assessments at the district level can be much more parsimoniously distributed. A district can get a representative sample of student knowledge by collecting data at critical points and intervals throughout the K-12 continuum. For example, if Algebra I is a critical course for the opportunity to learn the eleventh grade benchmarks, it would be logical to have all students take a district math assessment at the end of this course. The data from the various assessments for each content area, when viewed as a system, should provide a valid measure of the standards for that grade span (e.g., K-4).

Assessment types and levels: District assessments are those that are common throughout the district within specific grades or courses and should be administered and scored using uniform procedures. The district assessment system should help provide data for evaluating each school's improvement plan, but in cases where a school's improvement plan cannot be adequately evaluated using state or district assessments, a school will need to create some unique school-level assessments. These district assessments are often external to the classroom, but teachers should be able to use these data to plan and revise instruction and curriculum. Further, there is no reason why a well-developed classroom assessment or set of assessments cannot be "scaled-up" and used as the district assessment for a particular grade level and content area. In this case, there needs to be a method in place to ensure some commonality of assessment procedures and scoring across the district.

External assessments are often included in district assessment systems. These assessments can be standardized, which means that standard conditions of administration must be followed in all classrooms in which the tests are administered. This helps to assure comparability of results. Norm-referenced tests (NRT) and the PAWS are both examples of standardized assessments. District assessments are often standardized, as well. Large scale, standardized assessments are usually used to obtain information about curriculum and programs for school improvement planning and public reporting, although the results of these assessments can also be used with other data to help measure individual student achievement of the standards.

Many districts give norm-referenced tests as part of their assessment system. An "NRT" is different from the standards-based assessments given at classroom, district, and state levels. The major differences between these two types of assessments can be traced to the primary purposes of each type of test. With a norm-referenced test, students are simply compared to one another and to a representative "norm" sample of similar grade students throughout the country. Quality is judged in relation to other students and is often reported as percentile rank, which is the percentage of students that an individual outscored. Schools that give NRTs tend to do so because they wish to see how their students perform compared to a national sample of similar-aged students; whereas performance in a standards-based system is judged in relation to defined criteria of excellence. It is important to remember that most NRTs are not well-aligned with state standards, but if the district can evaluate the alignment, it can use the NRT to measure the standards where there is good coverage of the particular standard. When creating a district assessment system, the district should consider when norm-referenced assessments, PAWS, and other external assessments are administered so that particular grade levels are not burdened with an abundance of testing.

A Step-by-Step Approach [Note: The Wyoming #1 example is currently under revision]

Using the Results

Results of all levels of assessments—national, state, and local—should be used in thoughtful, responsible ways. By law, districts are required to provide a fair opportunity for all students to learn the standards. Disaggregation of assessment results helps to determine this has occurred and disaggregation is required for both accreditation and Title I. Schools are required to report student performance results to parents, schools, and communities. These results must be disaggregated by at least gender, ethnicity, and socio-economic status. Accreditation rules and guidance require schools to incorporate specific strategies and interventions into their school improvement plans to address subgroups that, by disaggregation of results, are not meeting district standards.

Schools and school districts could view these requirements merely as an obligation. However, it would be most beneficial to capitalize on the opportunity to use student performance

results as indicators of a need to improve school policies and programs, and to develop innovative instructional interventions that result in school improvement and improved student learning. This can also help a district identify potential remediation that may be necessary for specific groups of students and incorporate specific intervention programs to address these needs. With small districts and schools—which is true of most districts in Wyoming— it is necessary to look at several years and several sources of data prior to making any policy or programmatic decisions.

Fine-Tuning the System

Once districts start collecting data, it should be used to help fine-tune the system. You don't want to make huge changes that might affect the comparability of the assessments from one year to the next. But consider the following to help decide when to make some minor changes that will help the validity and reliability of the assessments.

➢ Are the data telling you what you expected? If the results don't make sense, reexamine the assessment. Is the assessment truly measuring the standards? For example, one district found that 99% of their first grade students were performing at the proficient or above level on the math assessment, with 95% of them in the advanced category. They knew this was not in keeping with their students' true performance in class. After reviewing the assessment, they determined that they had asked fairly low level questions and the standards were set too low.

➢ Are the tasks and tests so memorable that they lose meaning? If students remember the task so well that they can pass the information on to other students or can remember doing that task previously, then the results will not be true indicators about student knowledge of the standard.

➢ If you have more than one form of an assessment, make sure the assessment blueprints are the same. They need to have similar numbers of types of questions at equivalent levels of difficulty. They also need to maintain some of the same questions on both forms. Following these two suggestions helps to make sure that the multiple forms are equivalent.

Summary: Meeting the Accreditation Requirements

A district can design a technically adequate and useful assessment system by following the design principles outlined in Chapter II and the steps described in this chapter. The rules governing the measurement component of district accreditation can be found in Appendix B. As these rules (Appendix B) indicate, a district should have little trouble meeting the requirements for measurement in Chapter 6 if the suggestions in this document are incorporated into the design of district assessment systems. The required evidence to be submitted for accreditation as well as the accreditation criteria addressed by each source of evidence are summarized in Figure 11.

Required Accreditation Evidence	Criteria Addressed
District assessment plan that indicates: • Specific grade/course levels assessed, • Types of assessments, • Specific standards assessed, and • A brief description of the assessments. • Matrix of multiple measures in each common core area	Alignment, Multiple Measures
Evidence of alignment among the standards, benchmarks, and assessments	Alignment
Evidence of consistency of the assessment system	Consistency
Documentation regarding the fairness of the assessment system, including: • Procedures for ensuring the participation of all students regardless of disability or English language proficiency; • Participation rates for various subgroups of students including at least students with disabilities and limited • English Language Learners (ELL) for the various assessments in the system; and • A description of the methods used to include the most severely cognitively disabled students in the district assessment system	Fairness
A description of the standard-setting protocols used to establish cutscores for assessments in the system.	Standard Setting
Sample reports produced from the assessments in the system.	Fairness, Standard Setting, Multiple Measures
At least three sample assessments from the system that are not purchased, standardized assessments.	Alignment

Figure 11. Summary of Evidence required for District Accreditation.

CHAPTER IV: BODY OF EVIDENCE GUIDANCE

Introduction

The Wyoming State Legislature and the Wyoming State Board of Education have enacted high school graduation requirements for Wyoming students. The legislature and the State Board avoided a practice common to many other states of requiring students to pass a single graduation test. A major feature of the approach being implemented in Wyoming is the collection of evidence that will allow a school district to determine whether or not a student has met these new graduation requirements, rather than passing a single minimum competency test. Why is Wyoming using a Body of Evidence system and not a single test as many states are doing? There are many legitimate reasons for NOT using a single test to certify students for graduation, and there is no defensible educational reason for doing so. Following the recommendations of several major research organizations[6], we have chosen to rely on this locally determined set of data to base these important decisions. Most importantly, however, we believe implementing locally-designed assessment systems has the potential of leading powerful changes in instruction and learning in Wyoming's classrooms.

The purpose of this chapter is to provide guidance for the development of the components of the district assessment system necessary for evaluating whether students possess the knowledge and skills to meet the graduation requirements. Much of our thinking about a Body of Evidence assessment system evolved from conversations with many professionals in Wyoming school districts and the Wyoming Department of Education (WDE). The information in this chapter can be traced to a three-day meeting in June 2000 in Cody, Wyoming. The "Cody meeting" was the culmination of a series of meetings among many curriculum directors and WDE personnel and included more than 90 Wyoming school district representatives from 43

---

[6]American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1999). *Standards for educational psychological testing.* Washington, D.C.:American Psychological Association. Heubert, J. P. & Hauser, R. M. (eds.). (1999). *High Stakes: Testing for tracking, promotion, and graduation.* Washington, D.C.: National Academy Press.

Wyoming school districts. This document presents a framework for developing and implementing a Body of Evidence (BOE) assessment system. A single, "right" answer is not presented—because there is no single right approach.

This chapter begins with an overview of the working assumptions guiding the development of the Body of Evidence approach. Following this are the evaluation criteria by which district systems are be judged as well as the suggested evidence that districts can use to document that these criteria have been satisfied. These evaluation criteria are tied directly to the design principles discussed in Chapter II. A review of several types of "models" or approaches that could be used to meet the graduation requirements and fulfill the design principles is presented next, followed by a description of the evaluation process, including a discussion of the peer review teams. The chapter concludes with answers to commonly asked questions about Wyoming's graduation requirements and the Body of Evidence approach.

A Body of Evidence

What is a Body of Evidence? **A Body of Evidence, for our purposes, is an assessment system that is designed to determine whether students have met graduation standards and to provide a collection of evidence to support this decision.** While the concept of an assessment system was discussed earlier in this handbook (Chapter I), some key points bear repeating. In addition, issues specific to graduation requirements require further discussion. A well-designed Body of Evidence system will support inferences about performance related to the graduation standards where the information gleaned from the whole (the system) is greater than the sum of the individual assessments. A Body of Evidence graduation assessment system, therefore, has the following key characteristics (Coladarci et al., 2000):

> ➤ **The set of assessments comprehensively addresses the graduation standards.** A Body of Evidence system provides evidence about student achievement directly related to the Wyoming Content and Performance Standards.
> ➤ **The Body of Evidence assessment system provides students with multiple opportunities, using multiple formats, to demonstrate their knowledge and skills related to the Wyoming Content and Performance Standards.** In other words, the

system should allow *students to show what they know*, but the system should be designed in such a way so that *students who have not mastered the common core areas should not be "certified" as ready for graduation.*

➤ **Each of the individual assessments has a clear and explicit rationale.** When an assessment is included in the system, its purposes should be made clear and its connection to the rest of the system should be part of a logical and coherent plan. This is especially crucial in a Body of Evidence system where assessments can serve to certify student accomplishments across multiple standards and even multiple content areas. Without an explicit rationale for each assessment, the district can run the risk of having too many redundant assessments.

At the risk of oversimplifying, the Body of Evidence assessment system should be able to answer two basic questions:

> *1. Does the student know enough to graduate?*
> *2. Can the evidence support your answer to #1?*

Finally, the Body of Evidence system should be designed in accordance with the design principles described in Chapter II. If these principles are followed, each district's system should meet the criteria of an assessment system.

Working Assumptions Based on Rule and Statute

The following working assumptions were influential in the development of Wyoming's approach to designing a Body of Evidence assessment system. These assumptions are based on the legal requirements, defined by legislative mandates (W.S. 21-2-304) and State Board of Education decisions (Chapter 31 of Wyoming Education Rules).[7] For a more thorough review of these legal requirements, please refer to Appendix C.

---

[7] Chapter 31 Rules can be viewed at http://soswy.state.wy.us/RULES/5218.pdf

1. *The enabling legislation specifies quite clearly that all students receiving a Wyoming high school diploma must master the common core of knowledge and skills. In other words, all students are expected to meet the same standards and proficiency levels within each content area. The legislation maintains the single Wyoming high school diploma, but provides endorsements (advanced, comprehensive, or general), which shall be printed on a student's transcript.*

2. *While districts are responsible for establishing the meaning of the proficient performance level for the various subject areas, these levels must correlate closely with descriptions of proficient performance contained in the Wyoming Content and Performance Standards (see Section 8 of Chapter 31).*

3. *The Wyoming State Board of Education and the Wyoming Department of Education have established that Wyoming's framework for developing a Body of Evidence system relies on a compensatory approach for combining information at the benchmark and standard levels and a conjunctive approach at the content area levels. These various approaches for combining information are described in Appendix I.*

4. *The Wyoming Department of Education and the State Board of Education have adopted a mixed model for combining information within the Wyoming Body of Evidence System (please refer to Appendix J).* **When combining information at the benchmark and standards levels, districts must use at least a compensatory approach; and when combining information across content areas, a conjunctive approach must be used.** *This means that when determining whether or not a student has met a particular standard, it is the average (or some other compensatory approach) of the benchmarks that should be considered. In other words, students DO NOT have to demonstrate proficiency on every benchmark to be considered proficient on the standard. The same approach is true when combining information across standards to make a determination whether a student has demonstrated proficiency in the content area. Students DO NOT*

have to demonstrate proficiency on every standard to be considered proficient in the content area. On the other hand, students DO have to demonstrate proficiency in every content area as prescribed the requirements to earn one of three endorsements in order to receive a high school diploma. This proficiency requirement started with the graduating class of 2006.

Evaluation Criteria

The overarching goal of the district Body of Evidence system is that it should be able to support inferences about students' mastery of the common core content areas and the district's capacity for ensuring equality of educational opportunity. In order to meet this goal, specific design principles are necessary. **The design principles guide the development of districts' Body of Evidence systems in order to make sure the system fulfills the stated purposes**. The design principles described in Chapter II could be used as criteria for evaluating an assessment system, however, at the Cody meeting participants decided that the list of evaluation criteria should reflect the highest priority design principles. **Focusing on the highest priority design principles should lead to the development of solid district assessment systems, enabling district designers to concentrate their energy on the most important concepts**. In each of the subsections below, the evaluation criteria, written as target performance descriptors for districts attempting to meet these goals, are presented first (*italicized*). Following the principles of good assessment and evaluation, the evaluation criteria have undergone a series of revisions by peer review team leaders and WDE staff once actual documentation from districts was collected and peer evaluation teams were convened. In cases where the design principles presented in Chapter II need elaboration to better relate to designing a Body of Evidence system, additional information is presented; otherwise please refer to Chapter II. Finally, a description of the evidence that districts should submit to the peer review teams is presented at the end of each subsection (a summary of this required evidence is presented in Appendix L).

Alignment

> *The combination of assessments that comprise the system are aligned with district content and performance standards so that the full set of standards, both in terms of content and cognitive complexity, are assessed. If the district is incorporating course-based components (see pp. 51-52) in the Body of Evidence system, it must demonstrate that the curriculum is aligned with both the standards and assessments. In order to meet the alignment criterion, the district must indicate how the assessment system reflects the district's prioritization of the standards. Finally, multiple assessment measures and formats are employed in the system to maximize the alignment between standards and assessments.*

Recommended Alignment Evidence. The district's **assessment plan** is perhaps the most important piece of documentation to submit. When examining a district's BOE assessment plan a reviewer should be able ascertain the following:

➢ the major purpose(s) of the system,

➢ the specific grade/course levels assessed,

➢ the types of assessments employed,

➢ the specific standards assessed, and

➢ a brief description of the assessments.

Although a district could use a single assessment plan to indicate all of this information for all of the common core areas, it would probably be easier to have plans separated by content area with perhaps a macro look at the entire system.

Most alignment evidence will be evident in the district assessment plan, but the district needs to make sure that the processes used to ensure **alignment among the standards, benchmarks, and assessments are well documented**. This documentation should include assurances that the alignment was a two-way process between the assessments and standards and the assessment items/tasks were aligned to the levels of cognition called for in the performance

standards. If the district Body of Evidence system includes course-based information (e.g., grades), the alignment among the course curriculum, standards, and assessments needs to be evident and very well documented. In spite of all of the description and documentation, it is helpful to see actual **assessments** to fairly evaluate the system. The district should include a few illustrative assessments in each content area so that reviewers can develop a better flavor of the district's assessment system. Obviously, norm-referenced or other standardized tests should not be included since the review teams should be able to locate those easily.

Consistency

> *The decision regarding whether or not a student has met the graduation requirements for a given content area must demonstrate a high degree of consistency such that the misclassification rates are minimal.* *The focus of this evaluation should be concentrated on the system and should examine, for example, how different judges would evaluate the same set of data about a group of potential graduates. In order to satisfy this criterion, the district should also document that the results of the assessments are not overly influenced by error due to raters or the specific tasks/items used comprising the assessments. Individual assessments within the system need to be evaluated for consistency, in terms of error due to raters, tasks, administration conditions, and occasions.*

As mentioned in Chapter II, consistency or reliability is contingent upon the level of the decision being made and the amount of data (score points) gathered. For example, if the decision is whether or not a student is proficient in a content area, there will be a wealth of information and therefore, the decision most likely will be reliable. On the other hand, if the decision is whether or not a student has mastered a benchmark, there will be considerably less information available and likely less reliable. **This has major implications for those districts choosing to implement a conjunctive system at the standards level** (i.e., going beyond the state requirements for a compensatory approach at the standards level)—**the assessment system will be only as reliable as the least reliable assessment of each standard**. Obviously, the

ramifications are even more severe for those using a conjunctive approach at the benchmark level.

Recommended Consistency Evidence. There are many data sources that can contribute to this particular set of evidence including **traditional reliability evidence and/or inter-rater agreement** data from key assessments in the system. To judge the decision consistency associated with the system, **separate panels** can be convened to evaluate the evidence from the same set of students, especially those students close to the cutscores.

Fairness

The assessment system shall be designed, implemented, and evaluated so that it is not biased against any groups of students. Appropriate accommodations need to be employed so students with disabilities and English Language Learners have as fair a chance as possible to demonstrate what they know. Multiple assessment opportunities and formats should be used to maximize fairness. The results of the assessments comprising the system and the results of the system itself should be disaggregated to examine both the fairness of the assessment system and opportunities for all students to learn the standards.

Recommended Fairness Evidence. Evidence documenting the fairness of the system would include such things as the procedures used to ensure that the assessment system is not biased against any subgroups of students. The policies and procedures for ensuring the fair participation of all students in the assessment system regardless of disability or English language proficiency need to be included in the submission. Once the assessments are **pilot tested and data can be examined, districts will be required to disaggregate assessment** results by identified subgroups to search for potential bias in the assessment system and/or differences in opportunities to learn.

Standard Setting

The method for establishing cutscores between various performance levels on the Body of Evidence should be based on a defensible methodology and the district should

indicate a clear rationale for choosing one method over another. The method selected should incorporate clear descriptions of the performance levels and not based on arbitrary performance distinctions (e.g., traditional percentages).

Recommended Standard Setting Evidence. The district needs to provide documentation about the methods used to establish cutscores within the assessment system. In addition to describing the procedures, there should be a defensible rationale for choosing the particular approach. The district should provide evidence to document that cutscores derived from the standard-setting method used is valid in that students who are above the cutscore possess knowledge and skills related to the Wyoming Content and Performance Standard that students scoring below the cutscore do not.

Comparability

The assessments comprising the system should be comparable across schools and classrooms within the same school district both within a given year and across years. *The comparability requirement is, in some sense, a specific instance of the fairness criterion. If students' difficulty in meeting the graduation requirements is dependent on the particular school attended or the year they graduated, the system is not fair. The district should work to avoid changes in the difficulty of the assessments over time due to the "memorability" of specific assessments or changes in the way that scorers rate students' papers or projects.*

Recommended Comparability Evidence. The assessment plan can help satisfy the comparability criterion. The district should provide documentation that the assessment system requirements are comparable for students within each district both within a given class cohort and across years. Efforts to revise and replace assessment items/tasks should be documented. Finally, districts should attempt to provide evidence that the assessment requirements are at least on par with those in other Wyoming districts.

Secondary Design Principles

In addition to the primary design principles and evaluation criteria described above, there are several other considerations when designing a Body of Evidence assessment system. The following principles are still important to consider when designing the district system. Multiple measures, credibility, and consequences are all important considerations when designing a Body of Evidence system. Use of multiple measures is a particularly important principle because of the implications for improving alignment and fairness.

Putting It All Together—Building a System

Much of the information about designing a Body of Evidence system can be boiled down to a few main ideas (now that we have a common language) as districts work to continue to develop and fine-tune existing systems. Therefore, much of what follows is a review, but it also serves to synthesize the major ideas presented elsewhere in this document.

The first step in the design of any system is to **explicitly define the purpose(s)** of the system and to make sure those purposes are shared among key stakeholders. Concurrent with identifying the purposes, the district design team needs to **select a general approach to building a Body of Evidence system**. The choice of purposes and choice of model go hand-in-hand because certain models will be more likely to fulfill specific purposes. For example, if improving teachers' standards-based instructional and assessment skills is a primary goal of the system, then a system close to the classroom (e.g., course-based with common assessments) should probably be considered. Once the purposes are clearly defined, the district needs to explicitly **map the opportunities to learn the standards onto courses or other experiences**. This is especially critical if the district is including course-based approaches in its Body of Evidence system. In district-based systems, this mapping/alignment should also occur. While most districts have aligned their district standards with state standards, fully aligning the curriculum to standards and ensuring that students are provided an opportunity to learn all of the standards is ongoing in most districts. We cannot forget that students are also required to meet certain

Carnegie Units and districts should try, to the fullest extent possible, to mesh these two sets of requirements. If the opportunities to learn can be mapped onto courses that can fulfill Carnegie requirements, two sets of requirements can be satisfied, which should be easier to track and explain.

The next step is to **outline the primary system**—the way that most students will progress through the system. In other words, how will the typical student be provided the opportunity to learn the standards and to demonstrate that knowledge? For example, a system could be designed where most of the mathematics standards were embedded in Algebra I and Geometry courses, and this might be appropriate for most students. District officials could feel confident that students who successfully completed this course sequence had received a fair opportunity to meet the mathematics standards.

We know that many students will not have a fair chance to meet the graduation standards if they are expected to meet them in exactly the same way as all of their other classmates. Some students will need **alternate pathways to meet the graduation standards**. Continuing with the mathematics example, some students, by choice or prior preparation, might not be able meet the math standards through the algebra-geometry sequence. They might need to meet the mathematics requirements through applied or other types of mathematics courses. Ensuring that the appropriate standards are mapped onto non-primary approaches to meeting requirements should provide these types of alternate routes.

The last set of steps, which should go without saying at this point, is to **keep the design principles in mind as the system is developed and revised/updated**. Each of the various decisions should be weighed against each of **the highest priority design principles** to try to produce the most valid system possible. It is also important to recognize that certain decisions could help satisfy certain design principles while compromising others. For example, producing a homogeneous test could improve reliability, but alignment and validity would be compromised. In this case, alignment and validity should be weighted higher than reliability. **In general, alignment and fairness (two crucial components of validity) should be the highest priority design principles**, but like most aspects of assessment design, there are few clear rules about weighing these various principles to produce the best system possible. The following section describes some general approaches for designing a Body of Evidence system.

Possible Approaches for Designing a Body of Evidence System

There are several approaches districts can take to construct a Body of Evidence system and evaluate whether or not students have met the graduation requirements. The specific approach each district chooses should depend on the primary purposes identified by the district stakeholders. For example, if a district's primary purpose was the improvement of teaching and learning, then an approach relatively close to the classroom (e.g., some type of course-based approach) should be employed, whereas a district with accountability as the primary purpose might select a district-based approach. A range of possible approaches for meeting the Body of Evidence requirements are discussed below. The models are ordered by their proximity to everyday classroom practice. This is NOT a ranking of approaches—again, each district should select the approach that best matches its primary purposes. However, within each section, the potential advantages and disadvantages are described.

Course-Based Approaches

In a course-based approach, courses (or at least certain key courses) are clearly tied to standards; the grades in those courses are based on achievement of the standards, and then the Body of Evidence consists of course grades from these "standards" courses.

Potential advantages. Using this approach could increase teacher buy-in and could prove to be the most instructionally useful. In order for teachers to adjust instruction based on assessment data, the results need to be returned to teachers in close proximity to the assessment event so teachers can meaningfully adjust instruction. If teachers are responsible for designing and scoring the assessments, they could receive valuable feedback for modifying or targeting instruction. Teachers may also feel more respected if the Body of Evidence system was based solely on what happened in their classrooms and therefore, teachers might have a higher degree of "buy-in."

Potential disadvantages. There are several potential downfalls to this approach. The accuracy and consistency of classroom grading practices is a well-documented problem (Marzano, 2000), which could lead to credibility concerns with a district Body of Evidence system. Using a course-based approach would compromise comparability across the district if more than one teacher is responsible for teaching the standards in a particular content area. This is obviously not a problem in small districts with only one teacher in a given content area. This approach would also require more work and responsibility for individual teachers. This is the "flip-side" to the potential advantage regarding increased ownership by teachers—it also requires more work.

Course-Based & Common Assessment Approaches

This is similar to the course-based approach except that it includes the use of some common assessments to help provide some way to anchor grades from different teachers across the district. The common assessments would carry a certain percentage of the weight and information from the course would carry the remainder of the weight. The common assessments should carry enough weight (at least 25-30%) so they can realistically anchor the student grades. These common assessments DO NOT have to be "on-demand;" they can include common projects or other types of measures tied to standards. The work of the Body of Evidence Activities Consortium[8] is actually a special case of this approach. The Consortium has developed complex activities to assess multiple standards. Ideally, these activities could be imbedded in courses and serve as the common assessments.

Potential advantages. The major advantage of this approach is that it helps to ensure that students have equivalent opportunities to learn the standards, while still allowing a fair amount of individual teacher choice and discretion. This approach also encourages teacher collaboration because teachers have to come together to design and write the common assessments. If the common assessments were some of the activities designed by the Activities Consortium, teachers still have the opportunities for collaboration without having to do all of the work themselves.

[8] The Body of Evidence Activities Consortium/Wyoming Assessment Activities Consortium has been a cooperative arrangement among many Wyoming school districts, the Wyoming Department of Education, and the National Center for the Improvement of Educational Assessment. The Consortium's major purpose has been to design complex, integrated activities that can be used for both instruction and assessment and to provide professional development experiences for teachers and leaders so they developing complex activities to assess multiple standards. For more information about the Consortium, go to the web site at: http://www.wyaac.org.

Potential disadvantages. Depending on how much weight the common assessments received, there could still be potential problems as a result of teacher grading practices. Using these common assessments can help alleviate some of the grading problems because teachers' evaluations of student performance probably would tend to converge over time on the results of the common assessments, especially if the common assessments were perceived by the teachers as being high quality. The major disadvantage to this approach is that it requires time for teachers to collaborate. While collaboration should be encouraged, finding time to do so is a major concern.

District-Based Approaches

A district-based approach would rely on stand-alone assessments at key checkpoints in a student's high school career. These assessments could be at the end of specific courses or grades. The results from these assessments could comprise the Body of Evidence system as long as the set of assessments used were able to measure all of the standards appropriately as documented through the alignment evidence.

Potential advantages. One of the major advantages of this approach is simplicity. There are fewer records to keep and fewer decisions to make. It would also be easier to ensure technical quality as long as exams are updated regularly and if there were enough assessments in each content area to ensure reliability. This approach could have more credibility with the public, in part, because it would be fairly easy to explain.

Potential disadvantages. This approach could decrease teacher buy-in if they felt that students' performance in their class did not matter, rather only students' scores on the district assessments counted. It is doubtful that on-demand tests could cover the full range of standards and the district would need other types of assessments to measure all the standards (e.g., speaking/listening). As long as this was considered, it would not be a disadvantage. Additionally, the set of district assessments would have to represent a range of formats to satisfy the fairness and alignment criteria—using a single type of assessment would be a serious

disadvantage to this approach. Finally, if the same assessments are used over time, they will have to be updated (and equated to some extent) to ensure year-to-year comparability.

Multi-District Approaches

This is similar to the district-based approach but relies on collaboration among districts to reduce workload and to produce higher-quality assessments than can be produced by an individual district. This approach has many of the same advantages and disadvantages of the district-based approach, but also includes some unique features. This approach encourages/requires cross-district collaboration, which will help meet the Supreme Court's interest in "uniformity." Pooling the resources of many districts can help reduce the workload and improve the resulting products. The Body of Evidence Activities Consortium/Wyoming Assessment Activities Consortium discussed previously might fit into this category; but since the activities are designed to be course-based, it probably fits better into the course-based and common assessment approach discussed above. On the other hand, the Wyoming Assessment Collaborative formed in 2000 by the Region V BOCES would be characteristic of this approach. Districts using this approach would most likely need to establish common cutscores to help maintain credibility. It would be hard to explain why, for example, a score of 75% was NOT considered good enough in one district, but was good enough in another. The major disadvantage to this approach is that these relatively high stakes assessments would become memorable to students and teachers alike. There would need to be a plan to remove and replace items/tasks and equate the new forms to ensure year-to-year comparability. While it can be argued that this should be the case for any of the approaches, it is much more important when a district is relying on fewer, higher-stakes assessments.

Standardized Assessment Approaches

The results of standardized assessments such as norm-referenced tests (NRT), district achievement-level tests (DALT), and PAWS could be included in districts' Body of Evidence systems, but these assessments could **NOT constitute the entire Body of Evidence** system for any subject area. The major downfall for using any of these assessments is that, except for PAWS, most "off-the-shelf" assessments are NOT fully aligned with Wyoming Content and

Performance Standards. While PAWS is aligned with the mathematics and science standards, it is only designed to assess reading and writing in language arts—two of the three standards. While these tests can contribute important and technically sound information, too much reliance on them can have the unintended consequences of narrowing the curriculum so that teachers focus their instructional efforts on simply trying to ensure that their students perform well on these assessments. This could also be said for the multi-district assessments described above. Therefore, if these are used in the Body of Evidence system, they should only be considered one component in the set of multiple measures, and they should not receive enough weight to "drive the curriculum."

Mixed Models

It is easy to speculate that most districts' designs represent a combination of multiple\approaches described previously. A "mixed" model allows districts to rely on the best features of these multiple approaches thereby minimizing the disadvantages and maximizing the advantages. **It is also important to recognize that a district's Body of Evidence system does not have to look the same for each content area because some content areas might lend themselves to one approach over another**. The key is to be thoughtful and logical and to always keep the purposes in mind.

Evaluation Process

The evaluation criteria are only as good as the process by which they are judged. Of all the topics discussed at the initial Cody meeting, the proposed evaluation process generated the clearest consensus. Essentially the entire discussion that follows regarding the evaluation process was taken directly from discussion at the Cody meeting, except for a few modifications made necessary by State Board of Education actions. **In short, teams of peers convene to provide support and technical assistance to district leaders to help evaluate the district Body of Evidence systems**.

Peer Review Teams

There are two major purposes for using teams of peers to evaluate districts' Body of Evidence systems. First, involving practitioners helps to build assessment capacity in Wyoming school districts and helps distribute expertise across Wyoming. Second, involving peers in the evaluation process helps make the process fair and practical.

The peer review teams are comprised primarily of district personnel (e.g. superintendents, curriculum coordinators, special educators, principals, and teachers), Wyoming Department of Education staff, national experts, and at times other citizens. Team leaders are chosen each year of the review to lead the review teams. Both the team leaders and the reviewers receive training prior to conducting a review of a district's system. The training focuses on measurement principles as well as training to review district evidence to ensure consistency of judgments among team members. Many of the evaluation criteria have been and continue to be fine-tuned as peer teams comprised of district personnel review actual evidence from districts. Revisions to the Peer Review Rubric were made by WDE in collaboration with districts across the state during the 2007-2008 BOE work sessions.

The review teams do not accept or reject any BOE system. The review teams are responsible for providing feedback to improve or enhance a district's system based on the submitted evidence during the peer review process. The actual evaluation comes from the WDE through the district accreditation process. WDE uses the information from the peer review in district accreditation monitoring.

Peer Review Process and Schedule

Initial Review. As mandated by the State Board of Education Chapter 31 rules, Graduation Requirements, all Wyoming school districts were required to send documentation regarding their Body of Evidence systems to the Wyoming Department of Education during the 2001-2002 school year. The review team analyzed the submitted evidence and provided a report with suggestions for the district to improve its system.

The 2002-2003 Review. Districts resubmitted evidence and documentation supporting their Body of Evidence systems. The process was similar to the initial review except that an evaluation was filed with the accreditation unit of the Wyoming Department of Education, and

the reports were provided to the State Board of Education. Districts with Body of Evidence systems not meeting the specified criteria were required to make the necessary adjustments to improve the district's system.

      2003-2007. Districts file yearly updates to their Body of Evidence plans. Districts attended optional peer reviews offered by WDE in 2004, 2005, and 2006.

      The 2008 Review. Districts submit updated evidence of their Body of Evidence Systems at this optional review. Results will be reported to the State Board of Education, and subsequent technical support will be provided by WDE.

      The 2009 Review. Districts submit updated evidence of their Body of Evidence Systems at this required review. Results will be reported to the State Board of Education and will be tied to accreditation.

Some Questions and Answers About Wyoming's Graduation System

1. Why did the Wyoming Department of Education adopt a compensatory approach at the standard level for combining information instead of requiring students to demonstrate proficiency on all of the standards?

While WDE originally suggested that students demonstrate proficiency on all standards, this position has been modified as described above for several reasons. First, using a compensatory approach at the standard level recognizes that students have various achievement profiles that allow them to use their strengths to offset their weaknesses. For example, a student's exceptional performance in algebra can be used to offset weaker performance in geometry and still be considered proficient in mathematics. Second, research in other settings as well as research using WyCAS data allowed us to conclude that using a conjunctive approach at the standard level would be too stringent and could prevent too many students from graduating. Districts have the freedom to adopt a more stringent approach, such as using a mixed conjunctive-compensatory approach. In this type of approach, the district could determine that on average students need to

be above a certain cutpoint (compensatory), but not be any lower than a particular threshold level on any standard (conjunctive).

2. Does this mean that a district can "pick and choose" which standards and benchmarks to teach to students?

No, **ALL** students need to be provided an opportunity to learn **ALL** district standards and benchmarks. It would be very difficult for a student to "master" a content area without having an opportunity to learn all standards. For example, picture a high school course divided into eight units with eight assessments that contribute equally toward the final grade. If you only taught some students five of the eight units, what is the lowest average score on the five assessments they could have to pass the course (assume 70%=passing)? The answer to this little thought experiment is that the student could not score high enough on the five assessments to compensate for the zeros on the other three assessments. Even with scores of 100% on the five attempted standards, the student would have an average of only 62.5% and would not pass the class.

3. Can a district require students to demonstrate proficiency on all standards even though the state does not?

We have outlined the minimum requirements in terms of combining information. Districts cannot use a more lenient approach (e.g., disjunctive), but districts can require a more stringent approach such as a conjunctive approach at the standard level. We do not recommend that districts adopt a conjunctive approach at the standard level until after they have had a chance to implement and evaluate the impact of their systems. Research (e.g. Gribbons, 1999) and previous practice indicates that the compensatory approach is most appropriate and we believe it is also most aligned with the statutory requirements in W.S. 21-2-304.

4. In order to implement this system, does it mean that we (at the district level) have to keep track of student performance on every benchmark for every standard?

No, at the classroom level, teachers need to keep track of fine-grained information to make instructional decisions, but not all of this information needs to be collected at the district level. Districts need to maintain information about student performance relative to standards because that is what will be used to determine if students have demonstrated proficiency in the content area. If the set of assessments used to evaluate whether or not students have met the requirements for the content area are well-aligned to standards and benchmarks and this alignment is clearly documented (in terms of which questions, tasks, etc., map to which benchmarks), then it is not important to centrally maintain records regarding student performance on the benchmarks. Again, benchmark-level information should be useful to teachers and should be maintained for instructional purpose.[9]

5. But, there is still a lot of information—how can we keep track of it all?

It is important to bear in mind that the data collection demands are contingent on the type of Body of Evidence model employed by the district. There are many methods for keeping track of standard-level information, but using a truly compensatory system necessitates certain data requirements. For example, recoding assessment scores into "pass/fail" or "proficient/non-proficient" (i.e., 0, 1) discards potentially compensating information that can lead to misclassifying students. For example, imagine if your district used a decision-rule that students need to have an average score of 70% across the standards in a given content area to pass. For the purposes of this example, assume there were five standards and scores were provided on a four-point scale where 3 and 4 were "passing" and 1 and 2 were not. If a student had the following scores 4, 4, 4, 2, 2, they would be considered passing (16/20=80%), but if these data were reduced to pass/fail, the student would not pass (1, 1, 1, 0, 0= 3/5=60%). One caution, however, when using information in its "raw" form, multiple measures need to be placed on the same type of scale prior to combining test scores or other data.

6. How do we handle transfer students?

---

[9] Subsequent documents will present some concrete examples for (re)designing classroom grading systems. For now, the reader is urged to refer to Marzano (2000).

Many are concerned that the new graduation requirement will create problems when students transfer into a Wyoming high school. This concern manifests itself differently depending on whether the transfer students are coming from a Wyoming or an out-of-state school and whether the student transfers relatively early versus relatively late in their high school career. The one assumption being made here is that every Wyoming high school currently has at least one person for certifying transcripts to deal with transfer students. Dealing with Wyoming originating schools will be, in general, considerably easier compared with transfer students from other states. It is fair to assume that all Wyoming students, regardless of where they attend school are being provided an opportunity to learn the Wyoming Content and Performance Standards. Therefore, students coming in from another Wyoming school should be able to bring a record of the standards they have met as well as those standards they still need to meet. If there are any concerns, the receiving school should be able to call the sending school for clarification. Accepting students from out-of-state schools is clearly more challenging than the situation for in-state transfers. The challenge is progressively more difficult the further along a student is in her/his high school career. Registrars at receiving schools will have to do the best job possible to make decisions regarding a student's status relative to the Wyoming Content and Performance Standards. This might require calling the sending school to request copies of course syllabi and assessments or even talking with the principal or teachers. Receiving schools could even use some of their existing exams as "placement tests" to get a sense of the student's achievement of the Wyoming Content and Performance Standards.

7. How do we include students with disabilities in the Body of Evidence system?

Both federal and state laws required that all students with disabilities be included in general state and district-wide assessment programs, with accommodations, where necessary. Accommodations need to be provided for students without substantially altering the proficiency levels, the validity of the assessments, and/or the assessment system. There is a developing set of best practices for providing accommodations for students with disabilities. Here in Wyoming, the *Wyoming Accommodations Manual for Instruction and Assessment* provides guidance for using

accommodations for students with disabilities, or who are on a 504 Plan, or for English Language Learners.[10] These students should be accommodated, according to their Individual Education Plans (IEP) and the *Wyoming Accommodations Manual*, so that they are able to fairly demonstrate their achievement relative to the Wyoming Content and Performance Standards. However, this does not mean that students with disabilities are able to use "modified" assessments—those where the construct being measured is clearly altered. Further, all students in a given district are expected to meet the same performance levels in order to meet the graduation requirements.

8. Should PAWS be used in the Body of Evidence system?

Chapter 6, Section 8 of the Rules and Regulations states that among other measures, the district must incorporate the state assessment system into its district assessment system by using state assessment results to measure the Wyoming Content and Performance Standards in mathematics, reading, and writing. Though a close literal interpretation of the rule does not **require** the 11th grade PAWS to be used in a district BOE system, the PAWS must clearly be part of the district assessment system. Currently, PAWS provides information in relation to the mathematics, language arts, and sciences standards at the individual student level. A district may want to include PAWS scores as one component of their BOE system for those content areas. However, PAWS alone would not be sufficient by itself, and PAWS should never be used as the sole assessment of a standard, since many of the design criteria would not be met if this approach were used. WDE does not support systems where performance on PAWS is the sole reason used to deny a diploma.

---

[10] 12 In addition to the *Wyoming* manual, district personnel should visit the National Center for Educational Outcomes' (NCEO) website— http://cehd.umn.edu/nceo/ for information on assessing students with disabilities.

9. What about students who need remediation in order to meet the graduation requirements?

It is beyond the scope of this document to suggest remediation and extended learning strategies, rather this document is concerned only with the assessment aspects of identifying students who need remediation. The assessment system needs to be designed so that students who need remediation can be identified early in their high school career. As one participant at the Cody meeting put it, "there should be no surprises." Whether or not a district should implement a systematic screening process is a local decision, but it is necessary that districts try to ensure as early as possible that students will be on track to meet graduation requirements.

10. What are the implications for professional development as a result of implementing a Body of Evidence assessment system?

The professional development needs necessary for faithfully implementing a Body of Evidence system are continuous. Obviously, as the design of the Body of Evidence system gets closer to the classroom, the greater the professional development needs. The most pressing needs for teachers and principals are learning how to teach in a standards-based environment, including learning more about developing and scoring classroom assessments and using assessment data to make instructional decisions. Another major professional development need is learning to improve classroom grading practices. At the district level, there are professional development needs around the technical aspects of assessments as well as data analysis and interpretation.

Summary

This chapter is intended to provide guidance to district personnel as they work to design and implement assessment systems for determining whether or not students have met Wyoming's graduation requirements. This document should continue to be considered a work in progress. New legislation, rules, or State Board of Education decisions will undoubtedly influence how students are certified for graduation. Further, as more districts fully design and

implement Body of Evidence assessment systems, we will learn a great deal as we collect data and observe the consequences of various design decisions. Therefore this document will continue to be updated regularly in order to provide the most appropriate guidance to districts.

REFERENCES

Airasian, P. W. (1994). *Classroom assessment (2nd ed.).* New York: McGraw Hill.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1999). *Standards for educational psychological testing.* Washington, D.C.:American Psychological Association.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977).Frameworks for comprehending discourse. *American Educational Research Journal, 14,* 367-382.

Bohling, A. (1999, March). *Graduation requirements: A new focus on student learning in Wyoming.* Wyoming Department of Education. Cheyenne, WY.

Cizek, G. (2001) (Ed*.). Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Publishers.

Cohen, D. K. and Hill, H. (1998). *Instructional policy and classroom performance: The mathematics reform in California.* Ann Arbor, MI: University of Michigan.

Coladarci, T., Johnson, J. L., Beaudry, J., Cormier, M., Ervin, R., Rosenblum, J. M., and Silvernail, D. L. (2000). *Measured measures: Technical considerations for developing a local assessment system.* Augusta, ME: Maine Department of Education.

Elmore, R. F. and Rothman, R. (Eds.) (1999). *Teaching, testing, and searning: A guide for states and school districts.* Washington, DC: National Academy Press.

Gronlund, N., Linn, R., & Davis, K. (1999). *Measurement and assessment in teaching.* Prentice Hall.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000a). *Handbook for setting standards on performance assessments.* Washington, DC: Council of Chief State School Officers.

Herman, J. & Gribbons, B. (2000). *Using multiple measures to judge adequate yearly progress.* Final deliverable to the Wyoming Department of Education. Los Angeles, CA: Center for the Study of Evaluation, UCLA.

Heubert, J. P. & Hauser, R. M. (eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation.* Washington, D.C.: National Academy Press.

Jaeger, R. M. (1995). Setting standards for complex performances: An iterative judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, 14(*4), 16-20.

Kahl, S.R., Crockett, T.J., DePascale, C.A., & Rindfleisch, S.L. (1995, June). *Setting standards for performance levels using the student-based constructed-response method*. Paper presented at the annual meeting of the American Research Association, San Francisco, CA.

Kingston, N. M., Kahl, S.R., Sweeney, K., & Bay, L. (2000). *Using the Body of Work Method to Set Performance Standards: Experiences in Multiple States*. Measured Progress Research Report 2000-1. Dover, NH.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20, 8,* 15-21.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and evaluation in teaching (8th ed.).* New York: Macmillan.

Marion, S. F., Sheinker, A., Hansche, L., & Carlson, D. (1998, January). *Wyoming Comprehensive Assessment System design report to the Wyoming State Legislature*. Cheyenne, WY: Wyoming Department of Education.

Marzano, R. J. (2000). *Transforming classroom grading*. Alexandria, VA: Association for Supervision and Curriculum Development.

McTighe, J. & Ferrara, S. (1998). *Assessing learning in the classroom*. NEA.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational* measurement *(3rd ed.).* New York: American Council on Education, Macmillan Publishing

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Olson, L. (2001, January). Test debate: What counts as multiple? *Education Week,20*(16), 1. 10.

Plake, B. S. & Hambleton, R. K. (1998). *A standard setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work*. Paper presented at the annual meeting of the American Research Association, San Diego, CA.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know (2nd ed.).* Boston: Allyn & Bacon.

Proceedings of *Joint Conference on Standard Setting for Large-Scale Assessments*. Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O' Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction*. Boston: Kluwer Academic Publishers.

Rutherford, J. F. & Ahlgren, A. (1989). *Science for all Americans.* New York: Oxford University Press.

Salvia, J. & Ysseldyke, J. E. (1998). *Assessment (7th Ed.).* Boston: Houghton Mifflin.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 19,* 405-450.

Shepard, L.A.(1994). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In Aschbacher, P., Winters, L., & Herman, J. (1992). *A Practical guide to alternative assessment.* ASCD.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29, 7*, 4-14.

Shepard, L., Kagan, S. L., & Wurtz, E. (Eds.) (1998). *Principles and recommendations for early childhood assessments*. Submitted to the National Education Goals Panel by the Goal 1 Early Childhood Assessments Resource Group.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. Columbus, OH: Merrill.

Stiggins, R. (1998). *Classroom assessment for student success*. NEA.

Sweeney, K., Kahl, S. R., Kingston, N., & Bay, L. (2000, April). *Classroom Teacher Judgment and Body of Work Method: Wyoming Comprehensive Assessment System (WyCAS).* Paper presented at the annual meeting of the American Research Association, New Orleans, LA.

United States Department of Education. (1999, November). *Peer review guidance for evaluating evidence of Final Assessments Under Title I of the Elementary and Secondary Education Act*. Washington, DC: Author.

Webb, N. L. (1999, August). *Alignment of science and mathematics standards and assessments in four states:* Research Monograph No. 18. Washington, DC: Council of Chief State School Officers.

APPENDIX A
ONE STANDARD SETTING APPROACH

Since cutscores need to be derived in each of the nine common core content areas, we assume that this entire process needs to be repeated for each content area.

1. The first step in any standard-setting exercise is writing clear performance descriptors for the particular content and/or processes on which judgments need to be established. Representatives of the key stakeholders in the assessment system should by brought together to come to agreement about these performance descriptors. The importance of this step cannot be overstated. Without clear and agreed-upon descriptors, the rest of the process will falter.

2. The student work on the various assessments in the system should be scored and total scores (either weighted or not) should be calculated. If students can progress through the Body of Evidence system by taking very different sets of assessments, then this step and the remainder of the steps will need to be repeated for each "pathway" through the system. However, the descriptors written in the first step need to be held constant for each approach.

3. The next step involves arranging samples of scored student work along the score continuum. The actual number of student work samples necessary is dependent on the number of cutscores to be established and the range of student scores. To the extent possible, the panel should use fairly homogeneous (in terms of performance on the individual assessments) samples of student work. However, we recommend that the results of the standard setting (the cutscores) be tested against more heterogeneous collections of student work. This helps to make the process a little more straightforward. We recommend starting from the highest scoring student and working down the score continuum at regular intervals until you are satisfied that you have sampled students from

below the cut between the two lowest performance levels. For each cutscore (e.g., the score that separates proficient and advanced) we suggest trying to have 6-10 sets of student work that should span the cutscore. Using the proficient/advanced examples, it would be ideal to have work from several students who are clearly advanced and from several clearly proficient students and a slightly larger sample of work from students who might be considered high proficient and low advanced. The same principles apply to each cutscore. In the case of setting four performance levels, then three cutscores will be established.

4. A panel of at least eight people for each content area should be convened and trained. The first aspect of the training involves familiarizing the panel members with the performance descriptors. The second aspect of training would involve selecting a small training sample of student work, perhaps six sets of work along the score range. Panel members with the lead facilitator (a teacher or curriculum director who has a good understanding of the specific content area and the standard setting process) should then try to match each of the samples of work to the appropriate performance descriptor. Panel members would then discuss each rating and try to establish some clear understanding about the types of features in the student work they need to see as evidence to tie the work to the particular performance level.

5. After the facilitator and the panel feel confident that they understand the performance descriptors and the procedures for matching work to those descriptors, the panel will then repeat the process with the "live" sets of work. Each panel member should be provided with a rating sheet that lists the sets of work in the first column (by number) and has the number of additional columns for the number of performance levels being established as shown in the example below. Each panel member should individually rate each set of work according to the performance descriptor that it fits best. These ratings should be placed in the appropriate performance level column, as the "first rating" next to each student's number. Panel members should be encouraged to keep notes about each rating to help with the discussion described next.

Student #	First Rating				Second Rating				Notes
	Performance Level				Performance Level				
	Adv.	Prof.	Basic	Below Basic	Adv.	Prof.	Basic	Below Basic	
1									
2									
n									

6.   After panelists have finished their first ratings of all of the folders, the leader should first check agreement on a "master sheet" similar to the one pictured above and then facilitate discussions where there is not consensus. After the panelists finish the discussion, they should indicate on the "second rating" if they have changed their opinion. The facilitator should   then record the updated total ratings.

7.   When using this standard setting method with large-scale assessment programs, the next phase involves somewhat sophisticated statistical computations based on each panelist's rating of each folder of work. However, we think districts can approximate the statistical procedure and still arrive at much more meaningful cutscores compared with the fairly arbitrary demarcations typically used (e.g., 90% and above is "advanced"). The facilitator should   look   for   the   areas along the score continuum where the panelists have the highest degree of disagreement and try to hone in on a cutscore in that region. For example, imagine the following scenario:

a. 100% of the panelists agree that scores higher than 183 points should be considered advanced,

b. 60-70% of the panelists think that student scores of 175 match the advanced description,

c. 50% think that scores of approximately 171 match the advanced description while the other 50% think the work matches the proficient description, and

d. 30-40% think that scores of 167 match the advanced description while 60-70% think the work matches the proficient description.

In this scenario, we would expect the advanced/proficient cutscore to be in the 171 range. Please bear in mind that these score points are simply fictional for this example and unfortunately the data are rarely this clear-cut in real life standard setting. The same procedure and decision rules would apply when establishing the remaining two cutscores. These cutscores would then be used to place all of the individual bodies of evidence into the appropriate performance categories.

8.   Assuming the components of the Body of Evidence remain fairly constant, the same cutscores would be used in subsequent years. These cutscores would need to be adapted for students with slightly different BOE components. Instead of having to repeat this process for every student with different approaches, the district can use one of a few different solutions. One method would be to convert all the different scores to a common scale (e.g., percent of possible points) and simply fold these different collections of evidence into the process outlined in steps 2-7. Another approach would involve establishing the cutscores using the most common set of evidence and then convert the cutscores to a common scale and apply these cutscores to other collections of evidence. The key assumption in either of these cases—and this should be checked carefully—is that the collections of evidence should be equivalent in terms of alignment to the standards and the required cognitive demand. These approaches are technically "shaky," but are still more appropriate than setting arbitrary or traditional (e.g., 90%) cutscores.

APPENDIX B

WYOMING DEPARTMENT OF EDUCATION RULES AND REGULATIONS (2008): CHAPTER 6, SECTIONS 7 AND 8

CHAPTER 6: SCHOOL ACCREDITATION

Section 7. **Student Standards**.

All public school students shall meet the student content and performance standards at the level set by the school district aligned to state standards for:

(a) Common Core of Knowledge, emphasizing reading, writing and mathematics in grades one (1) through eight (8): (W.S. 21-9-101(b)(i) and (ii)).

(b) Common Core of Skills. (W.S. 21-9-101(b)(iii)).

(c) The district shall involve parents, community, and professional staff in developing student content and performance standards in the common core of knowledge and skills through an officially adopted planning process reinforced by board of trustee policies. Districts may choose to adopt state standards using the board of trustees' official process that includes involving parents, community, and professional staff. (W.S. 21-9- 101(b)).

(d) The district shall implement programs which will align to state standards and meet or exceed the requirements of those standards. Programs shall include:

(i) Planned strategies for intervening with students who fail to demonstrate proficiency on standards; and

(ii) Planned strategies for academically challenging students who consistently exceed standards.

(e) The district shall adopt and implement strategies to monitor the teaching of standards. (W.S. 21-3-110(a)(xvii), (xviii), and (xix)).

Section 8. **Student Assessment**.

(a) The state shall have a system of assessments aligned with state standards, consistent with the requirements of state and federal law. This assessment system shall be designed in accordance with standards of professional technical quality, as described in Section 8(f)(iii)(A) through (E), and be capable of generating results for all identifiable subgroups within each public school and public school district (W.S. 21-2-304(a)(iv) and (v)).

(b) For the 2004-2005 school year, each district shall ensure that all students in fourth, eighth, and eleventh grades participate in the Wyoming Comprehensive Assessment System (WyCAS).

(c) Beginning with the 2005-2006 school year, the state assessment system shall include and each district shall implement an assessment of student performance in reading, writing, and mathematics at grades three through eight and at grade 11.

(d) Beginning with the 2007-2008 school year, the state assessment system shall include and each district shall implement an assessment of student performance in science at least once in each of grades three through five, grades six through eight, and grades ten through twelve.

(e) The district shall ensure that all students enrolled in the grades required to be assessed participate in the assessment system in one of three ways (W.S. 21-2- 304(a)(v) and W.S. 21-3-110(a)(xxiv)):

 (i) In the general assessment with no accommodations;

 (ii) In the general assessment with appropriate accommodations; or

 (iii) In the alternate assessment.

(f) The district shall design and implement a district assessment system to measure student performance relative to district content and performance standards aligned to state standards. At a minimum, the district assessment system shall be designed to determine whether all students have had equality of educational opportunity to learn the content and skills represented in the standards and to the level established by the performance standards. This assessment system shall be designed in accordance with standards of professional technical quality, as set forth in Section 8(f)(iii)(A) through (E), and be capable of generating results for all identifiable subgroups within the district (W.S. 21-2-304(a)(iv) and (v), and W.S. 21-3-110(a)(xxiv)).

 (i) In order to evaluate equality of educational opportunity, the assessment system shall be implemented uniformly across the district.

 (ii) Among other measures, the district shall incorporate the state assessment system into its district assessment system by using state assessment results to measure the Wyoming Content and Performance Standards in mathematics, reading, and writing for fourth, eighth, and eleventh grade students.

 (iii) The system shall be designed and implemented so that inferences pertaining to equality of educational opportunity can be supported by the assessment system. The

system shall be designed to meet the following technical requirements, all of which contribute to documenting the validity of the overall district assessment system.

(A) Alignment-the combination of assessments that comprise the system shall be aligned with district content and performance standards so that the full set of standards in the common core of knowledge and skills, both in terms of content and cognitive complexity are assessed. The assessment system shall reflect ho w the district has prioritized the standards.

(B) Consistency-the assessment system should be designed and implemented in such a way so that inferences drawn from the results of the assessments are consistent and not dependent on error due to raters or the quality of the assessments. While the focus is on the system, in order to meet this requirement, individual assessments within the system will need to be designed to yield consistent results, in terms of error due to raters, tasks, administration conditions, and occasions.

(C) Fairness-the assessment system should be designed so that it is not biased against any group of students. As such, appropriate accommodations should be used so students with disabilities and Limited English Proficient students have fair access to the assessment system. As stated in Section 8(e)(i), (ii) and (iii) herein, multiple assessment formats should be employed in the assessment system which will contribute to improving the fairness of the system.

(D) In order to improve alignment, consistency, and fairness, multiple measures in each of the common core of knowledge and skill areas, but not necessarily at every grade level, shall be employed in the system.

(E) Descriptions of what constitutes proficient performance shall be clearly articulated and shall be correlated with the performance descriptors found in the Wyoming Content and Performance Standards. The cut scores that delineate the various performance levels on each assessment shall be tied to these district performance descriptors and shall be based on research or best practices.

(g) The district shall have a board-approved process in which student performance results are identified, monitored, and reported (W.S. 21-2-202(a)(xiv) and W.S. 21-2-304(a)(v)).

(i) The district shall distribute a uniform state report widely to its patrons in addition to other results from the district's assessment system as deemed appropriate by the district.

(ii) The district shall report the results of the district assessment system to the Wyoming Department of Education. Disaggregated results shall be reported for any of the following subgroups that include 10 or more students: gender, ethnicity, economic

status, mobility indicators, disability status, and other appropriate for the given locale. These disaggregated results shall be used to determine if all groups of students have been provided adequate opportunities to acquire the knowledge and skills necessary for meeting the graduation standards.

(h) In order for the State Board of Education to accredit school districts, each district will submit to the Wyoming Department of Education the following information at least 45 days prior to its scheduled accreditation visit:

(i) The district assessment plan that indicates the specific grade/course levels assessed, the types of assessments, the specific standards assessed, and a brief description of the assessments;

(ii) Evidence of alignment among the standards, benchmarks, and assessments;

(iii) Evidence of consistency of the assessment system;

(iv) Documentation regarding the fairness of the assessment system;

(v) Participation rates for various subgroups of students including at least students with disabilities and limited English proficient (LEP) students for the various assessments in the system;

(vi) Procedures for ensuring the participation of all students regardless of disability or English language proficiency;

(vii) A description of the methods used to include the most severely cognitively disabled students in the district assessment system;

(viii) Sample reports produced from the assessments in the system;

(ix) At least three sample assessments from the system which are not purchased, standardized assessments; and

(x) Other evidence that the district chooses to submit to support the technical quality of the assessment system.

APPENDIX C

WYOMING DEPARTMENT OF EDUCATION RULES AND REGULATIONS (2008): CHAPTER 31, SECTION 10

CHAPTER 31: GRADUATION REQUIREMENTS

Section 10. **Body of Evidence**

(a) Determination of proficient performance shall be demonstrated through a body of evidence identified by the district and approved by the district board of trustees. [W.S. 21-2-304 (a)(iii) and (iv)]. The body of evidence shall meet the following requirements:

The body of evidence assessment system shall be designed to best meet the needs of individual Wyoming school districts for certifying whether or not students have mastered the common core of knowledge and skills as embedded in the uniform student content and performance standards as specified in Section 8 of this chapter. The body of evidence assessment system shall be designed and evaluated according to the following criteria: alignment, consistency, fairness, standard-setting, and comparability.

(i) The alignment criterion shall be met if the combination of assessments that comprise the system are aligned with district content and performance standards so that the full set of standards, both in terms of content and cognitive complexity are assessed. Multiple assessment measures and formats shall be employed in the system to maximize the alignment between standards and assessments.

(ii) The decision regarding whether or not a student has met the graduation requirements for a given content area must demonstrate a high degree of consistency such that the rates of classifying students into performance categories incorrectly are minimal. The focus of this evaluation should be concentrated on the system and should examine, for example, how different judges would evaluate the same set of data about a group of potential graduates. In order to satisfy this criterion, the district should also document that the results of the assessments are not overly influenced by error due to raters or the specific tasks/items used comprising the assessments. Individual assessments within the system shall be evaluated for consistency, in terms of error due to raters, tasks, administration conditions, and occasions.

(iii) The body of evidence assessment system shall be designed, implemented, and evaluated so that it is not biased against any groups of students. Appropriate

accommodations shall be employed so students with disabilities and Limited English Proficient students have as fair a chance as possible to demonstrate what they know. Multiple assessment opportunities and formats shall be used to maximize fairness. The results of the assessments comprising the system and the results of the system itself shall be disaggregated to examine both the fairness of the assessment system and opportunities for all students to learn the standards.

(iv) The method for establishing cutscores between various performance levels on the district's body of evidence assessment system should be based on a research-based methodology and the district shall indicate a clear rationale for choosing their particular method. The method selected shall incorporate clear descriptions of the performance levels and should not be based on arbitrary performance distinctions (e.g., traditional percentages).

(v) The assessments comprising the system shall be comparable across schools and classrooms within the same school district both within a given year and across years.

(b) At a minimum, districts shall use a compensatory approach for combining assessment information at the benchmark and standard level when determining whether students have met the performance requirements for each common core content area.

(c) Districts shall use a conjunctive approach for combining assessment information across common core of knowledge and skills content areas to determine whether students have met the graduation requirements.

(d) A committee of peers shall review each district's body of evidence assessment system. The committee of peers shall recommend to the Superintendent of Public Instruction the district's status regarding its body of evidence assessment system. The committee of peers shall be comprised of Wyoming educators who have successfully completed peer review training conducted by the Wyoming Department of Education. The district shall submit evidence to the committee of peers in accordance with the peer review guidance provided by the Wyoming Department of Education based upon the evaluation criteria identified in Section 10 (a). This evidence shall include the following components: district assessment plans; evidence of alignment among standards, curriculum, and assessments; sample assessments; evidence of consistency, documentation of the standard setting methods, evidence supporting the fairness of the assessment system, documentation supporting the comparability of the assessment system across schools and years, and other documentation that the district chooses to submit to support the technical quality of the body of evidence assessment system.

(e) All Wyoming school districts with a high school shall submit their body of evidence assessment system documentation, as described in Section 10(d) of this chapter to the Wyoming Department of Education according to the following schedule:

(i) Districts shall submit body of evidence documentation by January 1, 2002, for a formative evaluation by the peer review teams and the Wyoming Department of Education. Written feedback regarding the quality of each district's body of evidence assessment system shall be provided, by June 15, 2002, to the district superintendent and board of trustees chairman.

(ii) Districts shall submit body of evidence documentation by January 1, 2003, for evaluation by the peer review teams and the Wyoming Department of Education. The State Board of Education, at the June 2003 meeting, shall incorporate the results of this review into each district's accreditation evaluation.

(iii) For the 2003-2004 school year and all following years, districts shall submit yearly updates to their body of evidence documentation to the Wyoming Department of Education. For the 2004-2005 school year and all following years, this documentation shall include the student performance results relative to the district's body of evidence assessment system including disaggregation of passing rates.

(f) The body of evidence for special needs students shall include accommodations in accordance with their individualized educational programs or 504 plans, and the policies as described in the *Policies for the Participation of All Students in District and Statewide Assessment and Accountability Systems,* which is available from the Wyoming Department of Education, 2300 Capitol Avenue, Hathaway Building, 2nd Floor, Cheyenne, Wyoming 82002-0050. These accommodations shall not substantially
alter the character of the assessments used to measure student performance.

APPENDIX D:

WYOMING SCHOOL DISTRICT #1 LANGUAGE ARTS ASSESSMENT MATRIX

[Note: This matrix is currently under revision by WDE]

APPENDIX E
DISTRICT ASSESSMENT PLAN SELF-ANALYSIS CHECKLIST

All areas for which there are standards are included in the assessment plan.

Multiple measures and formats are used to assess each standards area.

All standards are assessed.

Assessments are well-aligned to the standards.

Standards are well-sampled.

Assessments are given at logical checkpoints throughout the K-12 continuum.

The student population that is to be assessed and the time of year to give the assessment is
clearly stated for each assessment.

Participation policies and acceptable accommodations are clearly communicated.

Standards assessed by each assessment are indicated.

The purpose of each assessment and how the results will be used are clearly communicated.

APPENDIX F

SUMMARY OF STATUTORY AUTHORITY REGARDING GRADUATION REQUIREMENTS

<u>W.S. 21-2-304</u>

The legal history granting the statutory authority for the Wyoming State Board of Education to implement standards-based graduation requirements has been reviewed extensively in other documents (Bohling, 1999). For the purposes of this document, two subsections in the enabling legislation (§21-2-304(a)) are discussed. The key components of this statute, in terms of its relationship to the design of Body of Evidence assessment systems, are presented below. The statute is quoted directly and is then followed by the WDE interpretation and the implications for the design of the Body of Evidence assessment system.

> *(iii).… Student content and performance standards prescribed under this paragraph shall include standards for graduation from any high school within any school district of this state and shall described required performance levels in order to achieve proficiency of the common core of knowledge and common core of skills prescribed under W.S. §21-9-101(b).*

The common core of knowledge and skills relates to the broad content areas such as mathematics and language arts. However, in order to explicitly define these common core areas, the State Board of Education had to adopt uniform content and performance standards. Explicitly defining these common core areas–through standards and benchmarks—is critical for the design of curriculum and assessments. It would be virtually impossible to do so without this level of specificity. On the other hand, when making decisions about a student's graduation status, the judgment needs to be made at the more macro common core level rather than the individual standard level.[11]

---

[11] Obviously, student achievement information at the standard level is critical to making decisions about student performance in the content area.

The adoption of this particular stance has direct implications for the type of assessment system required. While the State Board of Education and Wyoming Department of Education have adopted this position (i.e., compensatory system at the standard level), this does not mean that districts can "pick and choose" which standards to teach and test. The requirement for providing equality of educational opportunity makes it clear that students need to be given an opportunity to learn *all* of the content and performance standards.

The next subsection of W.S. 21-2-304 states:

(a) (iv) Establish, in consultation with local school districts, requirements for students to earn a high school diploma as measured by each district's body of evidence assessment system prescribed by rule and regulation of the state board and required under W.S. 21-3-110(a)(xxii). A high school diploma shall provide for one (1) of the following endorsements which shall be stated on the transcript of each student:

(A) Advanced endorsement which requires a student to demonstrate advanced performance in a majority of the areas of the common core of knowledge and skills specified under W.S. 21-9-101(b) and proficient performance in the remaining areas of the specified common core knowledge and skills, as defined by the uniform student content and performance standards promulgated by the state board pursuant to paragraph (a)(iii) of this section;

(B) Comprehensive endorsement which requires a student to demonstrate proficient performance in all areas of the common core of knowledge and skills specified under W.S. 21-9-101(b) as defined by the uniform student content and performance standards promulgated by the state board pursuant to paragraph (a)(iii) of this section;

(C) General endorsement which requires a student to demonstrate proficient performance in a majority of the areas of the common core of knowledge and skills specified under W.S. 21-9-101(b) as defined by the uniform student content and performance standards promulgated by the state board pursuant to paragraph (a)(iii) of this section.

This section of the legislation provides the statutory authority for requiring some type of assessment system to determine if students possess the knowledge and skills necessary for graduation. The legislation clearly permits using tests to certify that students have mastered the common core of knowledge and skills, yet it also allows "other means" for demonstrating student mastery. The State Board of Education and the Wyoming Department of Education decided to require that districts collect a convincing array of evidence that can be used to determine whether or not a student has met the graduation requirements.

APPENDIX G

ANALYSIS OF WYOMING STATUTES DEFINING A PROPER EDUCATION FOR EACH WYOMING CHILD

§21-2-304. Duties of the state board of education.

(a) The state board of education shall:

*(iii) By rule and regulation and in consultation and coordination with local school districts, prescribe **uniform student performance standards** for the common core of knowledge and the common core of skills specified under W.S. §21-9-101(b) and promulgate **uniform standards for programs** addressing the special needs of student populations specified under W.S. §21-9-101(c). Student performance standards prescribed under this paragraph shall include standards for graduation from any high school within any school district of this state based upon performance or mastery of the common core of knowledge and common core of skills prescribed under W.S. §21-9-101(b). **Graduation standards imposed under this paragraph shall require the successful completion of the following components, as evidenced by passing grades or by the successful performance on competency-based equivalency examinations:***

> Student performance standards shall include **standards for graduation** based upon mastery of common core of knowledge and skills. These are the K-12 "Content Standards" that are further defined by the grade level benchmarks for each of the common core of knowledge and skills.

> The State Board of Education is required to prescribe uniform student performance standards for common core of knowledge and skills that define how well a student must perform AND uniform standards for **programs** which address the special needs of students defined in "special populations." Districts have formerly addressed the needs of gifted, LEP, and students with disabilities through programs that may look very different across the state. They need to be rigorous enough to support the student performance standards set by the State Board and to ensure uniformity among districts. WDE has always maintained that these children be held to the same standards as other children, however difficult it might be

(A)   Four (4) school years of English;

(B)   Three (3) school years of mathematics;

(C)   Three (3) school years of science; and

(D)   Three (3) school years of social studies, including history, American government and economic systems and institutions, provided business instructors may instruct classes on economic systems and institutions.

(iv) *Establish, in consultation with local school districts, requirements for students to earn a high school diploma as measured by each districts' body of evidence assessment system prescribed by rule and regulation of the state board and required under W.S. 21-3-110(a)(xxiv).* ***A high school diploma shall provide for one (1) of the following endorsements which shall be stated on the transcript of each student:***

(A) ***Advanced*** *endorsement, which requires a student to demonstrate advanced performance in a majority of the common core of knowledge and skills specified under W.S. 21-9-101(b) and proficient performance in the remaining areas of the specified common core of knowledge and skills, as defined by the uniform student content and performance standards promulgated by the state board pursuant to paragraph (a) (iii) of this section;*

(B) ***Comprehensive*** *endorsement which requires a student to demonstrate proficient performance in all areas of the common core of knowledge and skills specified under W.S. 21-9-101(b) as defined by the uniform student content and performance standards promulgated by the state board pursuant to paragraph (a) (iii) of this section;*

Graduation standards shall also include successful completion of specified number of school years of English, math, science, and social studies or passing competency-based equivalency exams.

Requirements for graduation are measured by each district's **Body of Evidence** assessment system.

Every student must master the common core of knowledge and skills in order to earn a high school diploma. One of three endorsements will be printed on the student's transcript: **advanced, comprehensive, or general.**

---

*(C) **General** endorsement which requires a student to demonstrate proficient performance in a majority of areas of the common core of knowledge and skills specified under W.S. 21-9-101(b) as defined by the uniform student content and performance standards promulgated by the state board pursuant to paragraph (a)(iii) of this section;*

APPENDIX H

STATE STANDARDS TIMELINE

On July 7, 2003, the Wyoming State Board of Education amended its Chapter 31 (Graduation Requirements) rules and adopted the revised state standards in all nine content areas and further establishing that the requirement of mastery of the state standards begin with the graduating class of 2006:

YEAR DEVELOPED	CURRICULAR AREA	YEAR IMPLEMENTED	YEAR REVISED
1997-98	Language Arts Mathematics	1998-99	**2003**
1998-99	Science Social Studies	1999-2000	**2003**
1999-2000	Foreign Language Health Physical Education	2000-2001	**2003**
2000-2001	Career/Vocational Education Fine/Performing Arts	2001-2002	**2003**

All areas listed in the Common Core of Knowledge and Skills are integrated into these nine groups of standards.

In addition to the state standards for the Common Core of Knowledge and Skills, there are specific requirements that instruction be provided at certain grade levels and in certain curricular areas:

"Not later than the 2002-2003 school year, all school districts shall provide instruction in foreign languages to students in kindergarten through grade 2 in accordance with standards promulgated by the state board." (W.S. 21-9-101(g))

"All schools and colleges in this state that are supported in any manner by public funds shall give instruction in the essentials of the United States constitution and the constitution of the state of Wyoming, including the study of and devotion to American institution and ideals, and no student shall receive a high school diploma, associate degree or baccalaureate degree without satisfactorily passing an examination on the principles of the constitution of the United States and the state of Wyoming. The instruction shall be given for at least three (3) years in kindergarten through grade eight (8) and for one (1) year each in the secondary and college grades." (W.S. 21-9-102)

APPENDIX I
WYOMING FRAMEWORK FOR STANDARDS

The State Board of Education through the Wyoming Department of Education utilized a framework for the development and implementation of uniform standards that allowed the districts as much flexibility as possible within their curriculum but yet would meet the requirement of uniformity. That framework included an integrated approach to the standards rather than course standards which would impose these courses upon the districts. This integrated approach provided the uniformity within districts' program standards but allowed them to fit the state standards into the courses of their own choosing within each curricular area. **The new approach also included developing standards that every student must master in order to earn a high school diploma, thus accomplishing both the requirement of uniformity and the requirement of setting graduation standards for the Common Core of Knowledge and Skills.** W.S. §21-2-304(a)(ii) requires that educational programs offered by public schools in accordance with the standards provide students an opportunity to acquire sufficient knowledge and skills, at a minimum, to enter the University of Wyoming and Wyoming community colleges, to prepare students for the job market or post-secondary vocational and technical training, and to achieve the general purposes of education that equip students for their roles as citizens. Therefore, the Wyoming standards are written by groups comprised of educators (both K-12 and higher education), parents, business people, and students. These groups are facilitated by the Director of Standards and the State Standards Coordinator of the Wyoming Department of Education. The development process for each set of standards is a year-long process during which drafts are written and comments are received and considered before the standards are presented to the Wyoming State Board of Education for approval.

The standards writing committees give careful consideration and thought as to **what every Wyoming child should know and do** in order to earn a Wyoming high school diploma. The K-12 (graduation) content standards are written first, followed by benchmarks at grades 4, 8, and 11 which further define the content standards. The drafts are then submitted to the public by providing a written copy to each school district as well as availability on the Wyoming

Department of Education web site at http://www.k12.wy.us. Comments are collected and reviewed throughout the period of development, with careful consideration to public input. In June of each year, the standards are submitted to the State Board of Education for its approval.

"Graduation standards" are the K-12 content standards. They define what students are expected to know and be able to do by the time they graduate. Students must master the K-12 graduation standards in order to earn a high school diploma. Districts track student performance relative to each content standard for graduation. This is done by tracking proficiency of the benchmark standards within each content standard and, for some skills, tracking proficiency at the content standard level.

"Benchmark standards" specify the content and skills that students must master along the way. The Wyoming state standards in all areas are written at grades 4, 8, and 11 benchmark levels. These benchmark levels help educators determine how students are progressing. They provide a target by which one can determine if a student needs remediation or enrichment. Districts have flexibility in the packaging of their courses and rolling the state content and benchmark standards into their curriculum and may have certain pieces of knowledge and skills which students learn in a grade level which is different from the state benchmark level, as long as the variation is reasonable. The expectation is that by the time students graduate from high school, they will have acquired the knowledge and skills set forth in the benchmark standards, which will provide a body of evidence that the content standards have been mastered. "Performance standards" describe how well students must perform the content and benchmark standards. These descriptors help teachers judge where students are performing in relation to the benchmark standards, and ultimately, the content standards. "Mastery" of the standards means that certain pieces of knowledge must be learned and certain skills must be acquired at a level of proficiency described in the student performance standards. Every student must reach that level of proficiency in order to master a certain standard. Mastery is achieved through proficiency over time based upon a body of evidence collected by the school district. Proficiency is a district decision. Each assessment used in the body of evidence to show mastery will have a level of proficiency determined for each specific assessment as aligned to the standards. Districts bear the responsibility to measure student learning, determine proficiency, and track student results for graduation using a body of evidence collected over time.

Body of Evidence

The body of evidence can be collected and tracked using a variety of systems. Standards can be tracked by passing courses aligned to standards, provided students' grades actually reflect student performance regardless of behavior, attendance, and effort. Proficiency can be collected on assessments aligned to the standards and tracked in individual portfolios or other tracking systems. Cut scores for determining proficiency on assessments must be established by the district. Benchmarks within a given K-12 content (graduation) standard can be weighted with an overall cut score for determination of mastery of that content (graduation) standard.

Requirements of Due Process for All Students and Implications for Students with Disabilities

Districts must adhere to the requirements of due process. There are primarily three considerations in the area of due process for the graduation requirements: opportunity to learn, adequacy of notice, and timely notice.

Opportunity to Learn

The standards must be taught to every child. A child cannot master that which s/he has never had the opportunity to learn. If a child is held accountable for material not taught him or her, this would violate the Equal Protection and Due Process clauses of the U.S. Constitution.

Adequacy of Notice

Distribution of circulars in the schools, individual mailings to some parents, and repeated announcements in the mass media may not be adequate notice. It may mean individualized notices to all parents.

Timely Notice

When changing the requirements to earn a high school diploma, more than two years, and perhaps from three to six years, are needed to allow a child sufficient time to gain the knowledge and skills required, depending upon the magnitude of the change in requirements.

Bullets of Information

• The Wyoming State Board of Education, in collaboration with the districts, sets the standards for graduation.

• Districts set the proficiency levels for all students, guided by the state performance level standards.

• The IEP team determines and designs instructional and assessment accommodations, including those related to the body of evidence.

• The IEP team does not set the proficiency levels nor substantially alter the standards. Mastery of standards for graduation will be demonstrated through a body of evidence collected by the districts. One of three endorsements will printed on a student's transcript: Advanced, Comprehensive, or General.

• The Wyoming State Board of Education has established 2006 for the first graduating class to show mastery of standards (K-12 content standards) to earn a high school diploma.

• A compensatory approach within each content area (meaning that proficiency on each and every standard is not required to demonstrate proficiency within a content area) has been established.

• A conjunctive approach is used across content areas (meaning that performance in each content area must be approached separately).

APPENDIX J
COMBINING ASSESSMENT INFORMATION

By definition, a Body of Evidence contains multiple sources of information about each student's academic achievement. Multiple assessments should be used in order to increase the validity of the inferences we can make about students' knowledge and skills. Determining the best methods for combining the various data sources in order to yield the most valid inferences possible is a major challenge facing district personnel and measurement professionals alike. The three general approaches (decision rules) for combining measures: conjunctive, compensatory, and disjunctive are presented below.

When using a **conjunctive** approach, scores on all measures used must be above the criterion point (cutscore) for the student to have met the overall standard. If three measures were used to determine whether or not a student has met a standard, the student would have to be above the cut score on each measure to be considered proficient. This is a fairly stringent approach and typically leads to the lowest pass rates.

A **compensatory** approach allows higher scores on some measures to offset (i.e., compensate for) lower scores on other measures. The most common example of the compensatory approach is the simple average. There are many more complex compensatory methods, such as many of the standard-setting processes used in large-scale assessment programs, but for now, thinking about the compensatory approach as a simple average will suffice.

A **disjunctive** approach is the most lenient, in that students only need to demonstrate proficiency on any one of the multiple measures used to be considered proficient. There can be legitimate reasons for using a disjunctive approach in certain settings, but it is a tough argument to make when there is political pressure to make sure that students meet high standards.

A **mixed-model** combines features of any or all of the three general approaches outlined here. For example, a model using both compensatory and conjunctive approaches might require students to score above a threshold score on every assessment (conjunctive), but above an average score on the full set of assessments (compensatory).

APPENDIX K

VALIDITY

Validity theory has been advanced considerably during the past thirty years. The recently Test Standards (APA, AERA, & NCME, 1999) confirmed that the construct (the theoretical and hypothetical trait we are intending to measure such as reading comprehension) is the unifying concept in validity evaluations. This means that validity of a test interpretation is more complex than simply assessing whether or not the items on a test match curriculum objectives or merely correlating the scores with some criterion measure. Evaluating the construct validity of interpretations from test scores would include the empirical evidence—content matching and test-criterion correlations—required of traditional validity investigations, the logical and theoretical attempts to delineate the construct, and the social consequences and values attributed to the test interpretation. The assumed connections among opportunities to learn, the ability of the system's measures to detect these opportunities, and the inferences about students from the set of measures are probably the most important line of theoretical inquiry.

Messick (1989) defined two important threats to construct validity. While these terms sound complex, they are quite straightforward and important to keep in mind when designing assessment systems. *Construct under-representation* means that inferences about student performance relative to the construct are based on measurement of only part of the construct. In other words, the construct could be under-sampled by the assessment items or the construct could be poorly defined in a theoretical sense and therefore under-represented on the assessment.

Construct irrelevance means that items on the assessment measure something other than, or in addition to, the focal construct. For example, a mathematics assessment with a very high reading demand is probably measuring reading in addition to mathematics. If reading achievement interferes with a student's ability to demonstrate what they know in relation to mathematics, then it could be argued that reading is at least partially irrelevant to the mathematics construct.

Following Cronbach, it is the socially constructed interpretation, not the test itself for which we are collecting validity evidence, because it is the interpretations that affect the decisions and consequences of a particular test and its use. Shepard (1993) advocates a more

straightforward means to prioritize validity questions. Using an evaluation framework, she proposes that validity studies be organized in response to the questions: "What does the testing practice claim to do?", "What are the arguments for and against the intended aims of the test [a values concern]?", and "What does the test do in the system other than what it claims, for good or bad?" (Shepard, 1993, p. 429). The questions are directed to concerns about the construct, relevance, interpretation, and social consequences respectively.

One line of evaluation, for instance, could focus solely on the technical quality of the assessments comprising the system. At its simplest, this would entail checking the match between the Wyoming Content and Performance Standards and the questions on the assessments (see the alignment section below). Because validity refers to the interpretations we make from test results, the procedures that are used to set performance standards and establish "cut scores" between performance categories would be part of a validity investigation (discussed in the standard-setting section). Another related line of research could examine the accuracy of these judgments. Every score contains a certain amount of error, and when making high stakes decisions, this error should be minimized as much as possible. Generalizability or reliability analyses that examine the consistency of scores (as well as trying to figure out the source(s) of the inconsistency) are important in the evaluation of technical accuracy.

A validity investigation would also focus on the consequences of the BOE assessment system. There are many consequences that are intended (e.g., improving academic achievement, equalizing opportunity-to-learn), and the system should be evaluated to see whether or not these intentions were fulfilled. There are often unintended consequences that result when an assessment system is added to the overall educational system. The investigation should be geared toward searching for and evaluating these unintended consequences of the assessment system. Some potential unintended consequences of this kind of graduation assessment program are misclassifying students (in either direction), a narrowing of the curriculum to focus only on tests, and the creation of an inordinate amount of work for teachers and other educators.

APPENDIX L

REQUIRED BODY OF EVIDENCE DOCUMENTATION

In order to ensure that the review process is as valid, fair, and reliable as possible districts are required to submit documentation so that the review teams can base their decisions on evidence and not on assumptions. Compiling the evidence using both the Peer Review Submission Guidelines and Peer Review Rubric allows district personnel a chance to evaluate their own system prior to submitting it to the peer teams.

BODY OF EVIDENCE SUBMISSION GUIDELINES AND RECOMMENDED DOCUMENTION FOR PEER REVIEW

General Structure of a BOE Plan:

In order for reviewers to easily locate the required documentation and evidence in each district's Body of Evidence Plan, districts may want to include the following:

- A cover that clearly identify your district binders
- Table of Contents
- Page numbers
- Appendices showing evidence to support each of the criteria

Section 1: Overview

The purpose of this section is to explain the overall Body of Evidence plan. After reading this section, reviewers should have an understanding of your district, and the approach taken.

For example, an overview may include the following:

☐ Demographics about the district (enrollment, # of high schools, etc.),

☐ Clearly define for students/parents the process by which a student graduates,

☐ Describe the type of system the district is using and the reasons for selecting that route (e.g., course based),

☐ An explanation of adjustments to the BOE system since implementation, and rationale for changes.

Submit two (2) content areas (one core and one non-core) that illustrate how the criteria are being implemented through the district's Body of Evidence. This will help reviewers see how all the criteria fit together into one coherent system for these two content areas.

Section 2: Alignment

In order to meet the alignment criterion on the BOE Peer Review Rubric, the submitted plan must include evidence of the following:

- There is documentation of **adequate sampling** of benchmarks as well as all the standards within the two representative content areas.

- There is evidence of a **two-way alignment process**: all assessment items and tasks align to standards and are represented in the assessments within the two representative content areas.

- The assessments from the representative content areas reflect the **cognitive depth** of the content standards and the types of student performance described in the performance standards.

Evidence in plan to support required criteria for alignment:

☐ Assessment samples for the representative content areas (1 core & 1 non-core) are included.
☐ Blueprints for the assessment samples are included in the plan.
☐ Matrices indicating all the assessments in the representative content areas (1 core & 1 non-core) and the standards and benchmarks assessed by each are included.
☐ The processes used by the district to ensure alignment of current standards and benchmarks as well as future changes are described.
☐ If the district Body of Evidence system includes course-based information (e.g., grades), the process for assuring alignment among the course curriculum, standards, assessments, and grading practices are described and appropriate polices included.
☐ Evidence of the processes used to ensure alignment of assessment items/tasks to the levels of cognition called for in the performance standards is present.
☐ Evidence of "think aloud" protocols and/or careful examination of student work is used to evaluate/document, and revise, if necessary, the alignment of its standards and assessment system.

Section 3: Consistency

In order to meet the consistency criterion on the BOE Peer Review Rubric, the submitted plan must include evidence of the following:

- For **open-ended assessments**, the district plan describes clear procedures to be used to ensure inter-rater reliability and defines a desired, acceptable rate. Data are presented that support implementation of the stated procedures.

- For **closed-ended assessments,** the district plan describes clear procedures to be used to ensure reliability and defines a desired, acceptable rate. Data are presented that support implementation of the stated procedures.

- If **teacher judgment** is part of the plan**,** the plan describes procedures to ensure reliability of judgment across assessments within a course and across teachers. There is clear documentation that judgment is anchored to the performance standards. Data are presented that support implementation of the stated procedures.

Evidence in plan to support required criteria for consistency:

☐ The procedures used to ensure inter-rater reliability on open-ended assessments are described.
☐ Inter-rater reliability data that meets acceptable rates (inter-rater reliability to meet or exceed 80% exact agreement and 98% exact + adjacent agreement) is included.
☐ The procedures used to ensure reliability on closed-ended assessments are described.
☐ Desired, acceptable rates of reliability on closed-ended assessments are stated.
☐ Reliability data on closed-ended assessments (to meet or exceed average reliability coefficients greater than 0.85) is included.
☐ Procedures used to ensure reliability of teacher judgment across assessments within a course and across multiple teachers are described.
☐ Reliability data of teacher judgment is included.

Section 4: Fairness

In order to meet the fairness criterion on the BOE Peer Review Rubric, the submitted plan must include evidence of the following:

- There is evidence the district uses procedures or tools to ensure that **assessment items/tasks** are not **biased** against subgroups of students.

- There is evidence the district uses **accommodations and alternate assessments** appropriately.

- There is evidence the district provides **multiple assessment opportunities**.

- A **variety of assessment formats and strategies** are included in the system.

- The district **disaggregates assessment results** (i.e. ethnicity, gender & socio-economic status) and the results are used to search for possible bias in the system.

- Relevant district data are presented to document that **participation rates** are at least 95% for all subgroups.

Evidence in plan to support required criteria for fairness:

☐ The procedures (e.g., bias committees) used to ensure that items and tasks are not biased against any subgroups of students are described.

☐ Sample forms and/or notes from bias review committee meetings are included.

☐ Policies and procedures for ensuring fair participation of all students in the system (e.g. students with disabilities or English language proficiency) are evident.

☐ There is evidence that illustrates accommodations and alternate assessments are used.

☐ There is evidence that the district system provides students with multiple opportunities, using different formats and strategies, to demonstrate their knowledge and skills.

☐ The plan includes disaggregated assessment results by identifiable subgroups (i.e. ethnicity, gender & socio-economic status) and describes how the district uses the information to make decisions.

☐ There is evidence that disaggregated assessment results are used to search for potential bias in the assessment system.

☐ The plan includes participation rates data for the content area assessments submitted.

Section 5: Standard-Setting

In order to meet the standard-setting criterion on the BOE Peer Review Rubric, the submitted plan must include evidence of the following:

• The district plan describes a **rationale** and a **defensible method** of standard-setting. It explains how the determination is made regarding proficiency levels in each content area.

• The plan identifies **cut scores** for each level of performance and the method used to determine these cut scores. It shows that they are clearly tied to performance standards.

• The district plan presents a timeline showing **adequate notification** to students on progress toward proficiency in each content area.

• There is evidence that the district has included **key stakeholders** (e.g., parents, community members, teachers) in the standard-setting process.

Evidence in plan to support required criteria for standard setting:

☐ The rationale and the standard-setting method used for determining proficiency at the content level are described.

☐ The cut scores used for each level of proficiency in the representative content area are included in the plan.

☐ The levels at which the cut scores have been set are clearly tied to the performance descriptors for the representative content areas.

☐ How and when individual scores are aggregated to make "graduate/not graduate" decisions are explained.

☐ The plan includes the timeline the district uses for their student notification process.
☐ The plan describes how key stakeholders are involved in the standard-setting process.

Section 6: Comparability

In order to meet the comparability criterion on the BOE Peer Review Rubric, the submitted plan must include evidence of the following:

- The district provides evidence that specific procedures are in place for ensuring comparability of assessments for all students **in a given year**, regardless of classroom, program, or school in the district.

- The district provides evidence that specific procedures are in place for ensuring comparability **across years.**

- The district provides evidence that specific procedures are in place for **replacing assessment tasks/items** with comparable tasks/items in terms of content, focus, and cognitive demand.

Evidence in plan to support required criteria for comparability:

☐ There is documentation of on-going district-wide trainings, common rubrics, the use of "seeded" papers, and common administration guidelines used to ensure comparability.
☐ The district has a process for ensuring the assessments are administered similarly from year-to-year.
☐ There is evidence that the district ensures that assessments are scored the same as in previous years (e.g., the use of anchor papers and common scoring rubrics, and scoring workshops for new teachers).
☐ The plan includes evidence of procedures for replacing assessment tasks/items such as the use of assessment blueprints and protocols.

APPENDIX M: BODY OF EVIDENCE PEER REVIEW RUBRIC

ALIGNMENT

☐ Meets Criteria (ALL bullets)	Evidence in Plan to Support Criteria	☐ Does Not Meet Criteria Highlight bullet(s) that apply	Recommendations
• There is documentation of **adequate sampling** of benchmarks as well as all the standards within the two representative content areas. • There is evidence of a **two-way alignment process**: all assessment items and tasks align to standards and are represented in the assessments within the two representative content areas. • The assessments from the representative content areas reflect the **cognitive depth** of the content standards and the types of student performance described in the performance standards.	☐ Assessment samples for the representative content areas (1 core & 1 non-core) are included. ☐ Blueprints for the assessment samples are included in the plan. ☐ Matrices indicating all the assessments in the representative content areas (1 core & 1 non-core) and the standards and benchmarks assessed by each are included. ☐ The processes used by the district to ensure alignment of current standards and benchmarks as well as future changes are described. ☐ If the district Body of Evidence system includes course-based information (e.g., grades), the process for assuring alignment among the course curriculum, standards, assessments, and grading practices are described and appropriate polices included. ☐ Evidence of the processes used to ensure alignment of assessment items/tasks to the levels of cognition called for in the performance standards is present. ☐ Evidence of "think aloud" protocols and/or careful examination of student work is used to evaluate/document, and revise, if necessary, the alignment of its standards and assessment system.	• The district provides little, incomplete, unclear or no evidence of **adequate sampling**. • The district provides little, incomplete, unclear or no evidence of **two-way alignment.** • The district provides little, incomplete, unclear or no evidence that the assessments reflect the **cognitive depth** of the content standards and the types of student performance described in the performance standards.	

CONSISTENCY

☐ Meets Criteria (First two bullets, and 3rd bullet only if teacher judgment is used in plan)	Evidence in Plan to Support Criteria	☐ Does Not Meet Criteria (Highlight bullets that apply)	Recommendations
• For **open-ended assessments**, the district plan describes clear procedures to be used to ensure inter-rater reliability and defines a desired, acceptable rate. Data are presented that support implementation of the stated procedures. • For **closed-ended assessments,** the district plan describes clear procedures to be used to ensure reliability and defines a desired, acceptable rate. Data are presented that support implementation of the stated procedures. • If **teacher judgment** is part of the plan**,** the plan describes procedures to ensure reliability of judgment across assessments within a course & across teachers. There is clear documentation that judgment is anchored to the performance standards. Data are presented that support implementation of the stated procedures.	☐ The procedures used to ensure inter-rater reliability on open-ended assessments are described. ☐ Inter-rater reliability data that meets acceptable rates (inter-rater reliability to meet or exceed 80% exact agreement and 98% exact + adjacent agreement) is included. ☐ The procedures used to ensure reliability on closed-ended assessments are described. ☐ Desired, acceptable rates of reliability on closed-ended assessments are stated. ☐ Reliability data on closed-ended assessments (to meet or exceed average reliability coefficients greater than 0.85) is included. ☐ Procedures used to ensure reliability of teacher judgment across assessments within a course and across multiple teachers are described. ☐ Reliability data of teacher judgment is included.	• The district provides little, incomplete, unclear or no evidence of procedures to be used to ensure inter-rater reliability on **open-ended assessments.** • The district provides little, incomplete, unclear or no evidence of procedures to be used to ensure reliability on **closed-ended assessments**. • The district provides little, incomplete, unclear or no evidence of procedures to be used to ensure reliability of **teacher judgment.** • The district provides little, incomplete, unclear or no evidence of desired, **acceptable rates of reliability** being defined. • The district provides little, incomplete, unclear or no evidence of **data** that supports implementation of the stated procedures.	

FAIRNESS

☐ **Meets Criteria** (ALL bullets)	**Evidence in Plan to Support Criteria**	☐ **Does Not Meet Criteria** (Highlight bullets that apply)	**Recommendations**
• There is evidence the district uses procedures or tools to ensure that **assessment items/tasks** are not **biased** against subgroups of students. • There is evidence the district uses **accommodations and alternate assessments** appropriately. • There is evidence the district provides **multiple assessment opportunities**. • A **variety of assessment formats and strategies** are included in the system. • The district **disaggregates assessment results** (i.e. ethnicity, gender & socio-economic status) and the results are used to search for possible bias in the system. • Relevant district data are presented to document that **participation rates** are at least 95% for all subgroups.	☐ The procedures (e.g., bias committees) used to ensure that items and tasks are not biased against any subgroups of students are described. ☐ Sample forms and/or notes from bias review committee meetings are included. ☐ Policies and procedures for ensuring fair participation of all students in the system (e.g. students with disabilities or English language proficiency) are evident. ☐ There is evidence that illustrates accommodations and alternate assessments are used. ☐ There is evidence that the district system provides students with multiple opportunities, using different formats and strategies, to demonstrate their knowledge and skills. ☐ The plan includes disaggregated assessment results by identifiable subgroups (i.e. ethnicity, gender & socio-economic status) and describes how the district uses the information to make decisions. ☐ There is evidence that disaggregated assessment results are used to search for potential bias in the assessment system. ☐ The plan includes participation rates data for the content area assessments submitted.	• The district provides little, incomplete, unclear or no evidence of plans, procedures, or tools to ensure that **assessment items/tasks** are not biased against any subgroups of students. • The district provides little, incomplete, unclear or no evidence that **accommodations and alternate assessments** are used appropriately. • The district provides little, incomplete, unclear or no evidence that **multiple assessment opportunities** are provided. • The district provides little, incomplete, unclear or no evidence that a **variety of assessment formats and strategies** are included in the system. • The district provides little, incomplete, unclear or no evidence of a process being used to **disaggregate assessment results** and the results are being used to search for possible bias in the system. • The district provides little, incomplete, unclear or no evidence that **participation rates** are at least 95% for all subgroups.	

STANDARD SETTING

☐ Meets Criteria (ALL bullets)	Evidence in Plan to Support Criteria	☐ Does Not Meet Criteria (Highlight bullets that apply)	Recommendations
• The district plan describes a **rationale** and a **defensible method** of standard-setting. It explains how the determination is made regarding proficiency levels in each content area. • The plan identifies **cut scores** for each level of performance and the method used to determine these cut scores. It shows that they are clearly tied to performance standards. • The district plan presents a timeline showing **adequate notification** to students on progress toward proficiency in each content area. • There is evidence that the district has included **key stakeholders** (e.g., parents, community members, teachers) in the standard-setting process.	☐ The rationale and the standard-setting method used for determining proficiency at the content level is described. ☐ The cut scores used for each level of proficiency in the representative content area are included in the plan. ☐ The levels at which the cut scores have been set are clearly tied to the performance descriptors for the representative content areas. ☐ How and when individual scores are aggregated to make "graduate/not graduate" decisions are explained. ☐ The plan includes the timeline the district uses for their student notification process. ☐ The plan describes how key stakeholders are involved in the standard-setting process.	• The district provides little, incomplete, unclear or no evidence of a **rationale** and a **defensible method** of standard-setting which describes how the determination of level of proficiency is made at the content level. • The district plan provides little, incomplete, unclear or no evidence of the **cut scores** used in each content area. • The district plan provides little, incomplete, unclear or no evidence that cut scores are clearly tied to performance standards. • The district plan provides little, incomplete, unclear or no evidence of a timeline showing **adequate notification** to students on progress toward proficiency in each content area. • The district plan provides little, incomplete, unclear or no evidence that **key stakeholders** have been involved in the standard-setting process.	

COMPARABILITY

☐ Meets Criteria (ALL bullets)	Evidence in Plan to Support Criteria	☐ Does Not Meet Criteria (Highlight bullets that apply)	Recommendations
• The district provides evidence that specific procedures are in place for ensuring comparability of assessments for all students **in a given year**, regardless of classroom, program, or school in the district. • The district provides evidence that specific procedures are in place for ensuring comparability **across years.** • The district provides evidence that specific procedures are in place for **replacing assessment tasks/items** with comparable tasks/items in terms of content, focus, and cognitive demand.	☐ There is documentation of on-going district-wide trainings, common rubrics, the use of "seeded" papers, and common administration guidelines used to ensure comparability. ☐ The district has a process for ensuring the assessments are administered similarly from year-to-year. ☐ There is evidence that the district ensures that assessments are scored the same as in previous years (e.g., the use of anchor papers and common scoring rubrics, and scoring workshops for new teachers). ☐ The plan includes evidence of procedures for replacing assessment tasks/items such as the use of assessment blueprints and protocols.	• The district provides little, incomplete, unclear or no evidence that specific procedures are in place for ensuring comparability of assessments for all students **in a given year**. • The district provides little, incomplete, unclear or no evidence that specific procedures are in place for ensuring comparability **across years.** • The district provides little, incomplete, unclear or no evidence that specific procedures are in place for **replacing assessment tasks/items** with comparable tasks/items in terms of content, focus, and cognitive demand.	