

Wyoming's Statewide Assessment System: Recommendations from the Wyoming Assessment Task Force

Compiled By

Joseph Martineau, Ph.D. and Scott Marion, Ph.D.
National Center for the Improvement in Educational Assessment

Draft September 25, 2015

October 2015

TABLE OF CONTENTS

Executive Summary	iii
Section 1: Appropriate Characteristics and Uses of Assessment.....	6
Introduction	6
Types of Assessments and Appropriate Uses	6
A Note on Classroom Assessment and Accountability.....	10
Section 2: Desired Characteristics and Uses	11
Introduction	11
A Statewide Summative Assessment or an Assessment System?.....	12
Section 3: Intended Outcomes	13
Introduction	13
Issues to Be Addressed.....	13
Intended Outcomes of a Comprehensive Assessment System	14
Section 4: Narrative Recommendations for a Comprehensive Assessment System.....	17
Introduction and Context	17
Proposed Wyoming Assessment System	18
Supports and Conditions.....	24
Ensuring a Successful Transition	25
Section 5: Detailed Recommendations	27
Introduction	27
Classroom Formative Assessment.....	27
District Assessment System	27
Interim Assessment.....	28
State Summative Assessment.....	30
Section 6: Potential Qualifying Products.....	40
Language Arts and Mathematics	40
Science.....	41
Section 7: Recommendations for Policy Coherence.....	43
References.....	46
Appendix A: Understanding Formative Assessment	47
Definition of Formative Assessment	47
Vignettes of Formative Assessment in Practice.....	49
Appendix B: One-Page Summary of Formative, Interim, and Summative Assessment.....	51
Appendix C: Detailed Highest Priority Uses and Characteristics.....	52
Appendix D: Mini-summative vs. Modular Interim Assessment Designs	54
Appendix E: Matrix Sampling to Reduce Required State Testing Time	59

EXECUTIVE SUMMARY

The Wyoming Assessment Task Force was formed to develop recommendations for a new Wyoming state assessment program or system. Those recommendations are presented here. The Task Force considered issues with the existing state and district assessment systems, and developed intended uses (Section 2) and intended outcomes (Section 3) for a new Wyoming comprehensive assessment system to address their concerns.

The Task Force identified the following critical issues:

- The general incoherence of results coming from the myriad of assessments.
- The amount of testing time required by the combination of state summative assessments, interim assessments, and district assessment systems.
- The general absence of timely, instructionally and programmatically useful results from the myriad of assessments.
- Confusion about the content standards that should be taught in Wyoming schools.
- The quality of district assessment systems and the level of assessment literacy among Wyoming educators.

Incoherence among Various Assessments

To address this concern, the Task Force recommended that the state-provided interim assessment be tied directly to the state summative assessment, and that it include the same format and types of tasks as included in the summative assessment. This would be accomplished by requiring vendors to bid on an interim assessment tied directly to the summative assessment.

To address issues of coherence between district assessment systems and state-provided assessments, the Task Force recommended that the district assessment systems be built to the same content standards as the state assessments, using the same formats and types of test questions.

Amount of Testing Time

To reduce the amount of testing time required of students, the Task Force recommended that use of the interim assessment no longer be required. It recommended that the state interim assessment be provided as a service to districts desiring to adopt it as part of their district assessment systems. It further recommended substantial flexibility for districts in the timing and manner of using state-provided interim assessments for districts that adopt them to maximize the usefulness of the interim assessment for each district. Because the state-provided interim assessment will be tied directly to the state summative assessment and to the official Wyoming state standards, districts will have a considerable incentive to adopt it.

The Task Force further recommended that a strict limit be placed on the amount of time that may be devoted to responding to state summative assessments no more than *one percent* of the required instructional hours for a given grade level. This limit provides enough time to allow for high-quality assessment of complex knowledge and skills *and* to restrict testing time to a reasonable level.

Finally, the Task Force recommended that the Department of Education work with a group of stakeholders to provide flexibility in the amount of time devoted to each testing session in the summative assessment to help schools and districts minimize disruptions to daily instruction.

The Need for Timely Results Useful for Important Educational Decisions

The Task Force indicated that a balance must be struck between scheduling the state summative assessment as late as possible in the instructional year and returning the results in time for use in school improvement activities, such as evaluating and adjusting interventions, curriculum, and programming during the late summer months. This, in essence, requires giving the test later and getting the results back sooner (a difficult task).

To address this need, the Task Force recommended that the assessment be moved fully online to expedite scoring and the return of results. It recommended that the test be moved closer to the end of the school year, and that the results be returned by the beginning of August each year when educators typically return for school improvement activities.

The Task Force recognized past problems with online assessment in the state, and provided a comprehensive set of recommendations to ensure a smooth transition. Key among these recommendations is that schools, districts, and the state be given until the spring of 2018 to implement the new state summative assessment. Without this lead time, a successful, smooth transition will be unlikely.

The Task Force also indicated a need for balancing the use of complex items types on the interim assessment with the need for near immediate reporting. To address this need, the Task Force recommended that reporting on the interim assessment can take up to one week from a student completing the interim assessment to assure that the results remain relevant to instruction. To make this feasible, the Task Force recommended that any complex item types that preclude reporting within a week of test completion not be included on the interim assessment.

Finally, the Task Force recognized the importance of high-quality, transparent reports in making educational decisions. It recommended that a rigorous report development process be implemented to target reports to the various audiences (e.g., students, parents, teachers, administrators, policymakers, and the general public) of state assessment to address each audience's critical needs while minimizing the possibility of misinterpretation. To improve accessibility of reports, the Task Force also recommended that a high-quality, easily navigable, dynamic reporting system be developed to house the reports for each audience. To serve the same purpose, the Task Force also recommended that state data systems be enhanced to ensure that each individual educator may obtain access to reports only for students he or she is currently responsible for, and to track the group progress of students he or she was previously responsible for.

Achieving Clarity on the Content to be Learned and Taught

The Task Force indicated that the sole use of the ACT for high school accountability has caused confusion about which standards will be taught in Wyoming high schools: the ACT college readiness standards or the official Wyoming state standards.

To address this confusion, and to provide more freedom to Wyoming's high school Juniors and Seniors, the Task Force recommended creating a split between assessment in grades 3-10 and grades 11-12. The Task Force recommended that assessments built to measure the Wyoming state standards be administered in grades 3-10, but not beyond, and that the grade 10 test be added to the criteria for Hathaway scholarship eligibility. In order to maintain the benefits of a college entrance examination, the Task Force further recommended that grades 11 and 12 be reserved for college entrance, work skills, Advanced Placement, International Baccalaureate, and other specialized testing. It also recommended that to better match individual students' interests, each 11th grader be required to take *either* a college entrance assessment or a work skills assessment.

The Task Force also indicated that the restriction of state summative assessments to multiple choice questions has also caused confusion because it is difficult to measure the complex knowledge and skills described in the Wyoming state standards with just multiple choice questions.

To address this confusion, the Task Force recommended that any type of test question appropriate to measure the Wyoming state standards be used on state-provided assessments, so long as time limits on state testing can still be met when including such complex question types.

Finally, the Task Force recommended the inclusion of Writing in the Language Arts assessment to signal that the state standards on writing are important, and to improve both the learning and instruction of writing.

Improving the District Assessment Systems and Assessment Literacy

The Task Force identified improving the assessment literacy and knowledge of appropriate data use for educational decision-making as critical to improving the quality of district assessment systems. The Task Force made several key recommendations to address these issues, including the following:

- Inclusion of a rigorous review of district assessment systems in accreditation
- The state hiring or contracting with an expert in interim and summative assessment to be a consistent presence on accreditation visits
- The provision of high-quality formative feedback to districts from the assessment expert to help them improve their systems
- The state supporting (but not directing) local districts or consortia of districts in providing professional learning activities to both teachers and administrators around classroom and district assessment.

Finally, The Task Force's recommendation to tie interim assessments directly to the state summative assessment is also likely to improve the quality of district assessment because of the resources that can be brought to bear on a state-provided assessment.

SECTION 1: APPROPRIATE CHARACTERISTICS AND USES OF ASSESSMENT

Introduction

In 2015, the Wyoming Legislature passed Enrolled Act 87, authorizing the State Board of Education to evaluate Wyoming's current state assessment system and the creation of the Wyoming Assessment Task Force. Specifically, Section 6 of the act authorizes:

The state board shall assemble a task force to assist with the assessment review and evaluation. The task force shall be comprised of representatives of small and large school districts and schools from all geographic regions of the state and shall at minimum include representatives from district and school administration, school district assessment and curriculum program administrators, elementary and secondary school teachers, school district board members, state higher education representatives, member of the Wyoming business community and parents of children enrolled in Wyoming public schools.

The twenty-four task force members¹ met seven times between June 1 and October 1, 2015. Three of these meetings were held in person, one of which was for two full days, and the remaining four meetings were held as webinars. This report presents the results of the task force deliberations. Before moving to the discussion of the task force recommendations, we first present in this section of the report some critical definitions and background assessment information on the appropriate characteristics and use of assessment.

We begin by defining two broad categories of assessment use: (1) *high-stakes accountability uses* and (2) *lower-stakes instructional uses*. Stakes may be high for students, teachers or administrators, or schools and districts. For students, test scores may be used for making high-stakes decisions regarding grades, grade promotion, ability grouping, graduation, admission to postsecondary education or training, and scholarships. For educators, student test scores may formally or informally factor into periodic evaluations used to inform important employment decisions in classrooms, departments, schools and districts. In addition, students, teachers and administrators are affected by high-stakes uses of test scores in school and district accountability: identification as a school or district in need of intervention often leads to involuntary interventions intended to correct poor outcomes.

Lower-stakes instructional uses of test scores for teachers and administrators include informing moment-to-moment instruction; self-evaluation in teaching a unit and adjusting subsequent plans accordingly, evaluating one's own instructional effectiveness; and evaluating the success of a curriculum, program, or intervention.

As described above, within the *high stakes accountability* and *lower stakes formative* categories of use, there are many potential uses. The multiple appropriate uses of the various types of assessment introduced below may fall into both broad categories.

Types of Assessments and Appropriate Uses

¹ There were 26 original members, but two members resigned during the course of the project due to other commitments.

While there are several possible categorizations of assessment by type, this section of the report reviews only one particularly relevant to the work of the Task Force: the distinction among *summative*, *interim*, and *formative* assessment². In this report, the three types of assessment are always discussed in this order *except for* defining them below. They are defined below in the order formative, summative, and interim because interim assessment is defined in relation to both formative and summative assessment.

This section provides definitions of the three types of assessment and outlines the appropriate uses of data gathered from them. Appropriate uses are underlined for emphasis. These definitions are critical to understanding what each type of assessment can and cannot do. Appendix B provides an at-a-glance summary of the typical characteristics, appropriate uses, and examples of each type of assessment.

Formative Assessment

Formative assessment has also been called formative instruction. The purpose of formative assessment is to evaluate student understanding against key learning targets, provide targeted feedback to students, and adjust instruction on a moment-to-moment basis.

In 2006, the Council of Chief State School Officers (CCSSO) and experts on formative assessment developed a widely cited definition (Wiley, 2008):

Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievements of intended instructional outcomes (p. 3).

The core of the formative assessment process is that it takes place during instruction (i.e., “in the moment”) and under full control of the teacher to support student learning while it is developing. This is done through diagnosing on a very frequent basis where students are in their progress toward learning goals, where gaps in knowledge and skill exist, and how to help students close those gaps. Instruction is not paused when teachers engage in formative assessment. Formative assessment covers fine-grained learning targets that are often the focus of a single unit of instruction.

Formative assessment is not a product, but an instruction-embedded process tailored to monitoring the learning of and providing frequent targeted feedback³ to individual students. Effective formative assessment occurs frequently, covering small units of instruction (such as part of a class period). If tasks are presented, they may be targeted to individual students or groups. There is a strong view among some scholars that because formative assessment is tailored to a classroom and to individual students that results cannot be meaningfully aggregated or compared.

Data gathered through formative assessment have limited to no use for evaluation or accountability purposes such as student grades, educator accountability, school/district accountability, or even public reporting that could allow for inappropriate comparisons. There are at least four reasons for this: (1) if carried out appropriately, the data gathered from one unit, teacher, moment, or student

² In defining formative, interim, and summative assessment, this section borrows from three sources (Perie, Marion, Gong, & Wurtzel, 2007; Michigan Department of Education, 2013; Wiley, 2008).

³ See Sadler (1989).

will not be comparable to the next; (2) students will be unlikely to participate as fully, openly, and honestly in the process if they know they are being evaluated by their teachers or peers on the basis of their responses; (3) for the same reasons, educators will be unlikely to participate as fully, openly, and honestly in the process; and (4) the nature of the formative assessment process is likely to shift in such a way that it can no longer optimally inform instruction.

Because there is considerable confusion about what formative assessment is, further definition and four vignettes⁴ describing formative assessment in action are given in Appendix A to clarify the meaning using concrete ideas. The first two vignettes are also presented in condensed form in the one-page summary of formative, interim, and summative assessment in Appendix B.

Summative Assessment

Summative assessments are generally infrequent (e.g., administered only once to any given student) and cover major units of instruction such as semesters, courses, credits, or grade levels. They are typically given at the end of a defined period to evaluate students' performance against a set of learning targets for the instructional period. The prototypical assessment conjured by the term "summative assessments" is given in a standardized manner statewide (but can also be given nationally or districtwide) and is typically used for accountability or to otherwise inform policy. Such summative assessments are typically the least flexible of the various assessment types. Summative assessments are also used for testing out of a course, diploma endorsement, graduation, high school equivalency, and college entrance. Appropriate uses of such standardized summative assessments include school accountability, district accountability, curriculum evaluation, program evaluation, and informing policy-makers in high-level decision-making. Depending on their alignment to classroom instruction and the timing of the administration and results, they may also be appropriate for grading.

Less standardized, but no less summative, assessments are also found in the majority of middle- and high-school classrooms. Such assessments are typically completed near the end of a semester, credit, course, or grade level. Common examples are broad exams or projects intended to give a summary of student achievement of marking period objectives, and figure heavily in student grading. Such assessments tend to be labeled "mid-terms," "final projects," "final papers," or "final exams" in middle and high school grades. Elementary school classrooms also have similar summative assessments but these do not have a consistent label in elementary grades. Classroom summative assessments may be created by individual teachers or by staff from one or more schools or districts working together.

Summative assessments tend to require a pause in instruction for test administration. They may be controlled by a single teacher (for assessments unique to the classroom), groups of teachers working together, a school (e.g., for all sections of a given course or credit), a district (to standardize across schools), a group of districts working together, a state, a group of states, or a test vendor. The level at which test results are comparable depends on who controls the assessment. They may be comparable within a classroom, across a few classrooms, within a school, within a district, across a few districts, within a state, or across multiple states.

⁴ Informed by Wiley (2008).

Appropriate uses of such summative assessments include student grading in the specific courses for which they were developed. If designed well, they can also be used to adjust curriculum, programming, and instruction the next time the large unit of instruction is taught; and to serve as a post-test measure of student learning. If the assessments are well-designed and a carefully- and well-defined set of rules is in place for appropriate administration, scoring, and use of results they may also be reasonably used for accountability.

Interim Assessment

Many periodic standardized assessment products currently in use that are marketed (or otherwise labeled) as “formative,” “benchmark,” “diagnostic,” or “predictive” actually belong in the interim assessment category. They are neither formative (they do not facilitate moment-to-moment targeted analysis of student learning, frequent feedback to students and teachers, or timely adjustment of instruction) nor summative (they are not intended to provide a broad summary of achievement of course- or grade-level learning objectives tied to specific state content standards). In contrast to formative assessment

Many interim assessments are commercial products and rely on fairly standardized administration procedures that provide information relative to a specific set of learning targets—although not always tied to specific state content standards—and are designed to inform decisions at the classroom, school, and/or district level. In some cases, interim assessments may be controlled at the classroom level to provide information for the teacher, but unlike formative assessment, the results of interim assessments can be meaningfully aggregated and reported at a broader level. However, the adoption and timing of such interim assessments are likely to be controlled by the school district. The content and format of interim assessments is also very likely to be controlled by the test developer. Therefore, these assessments are considerably less instructionally-relevant than formative assessments in that decisions at the classroom level tend to be *ex post facto* regarding post-unit remediation needs and adjustment of instruction the next time the unit is taught.

Common assessments developed by a school or district for the purpose of measuring student achievement multiple times throughout a year may be considered interim assessments. These may include common mid-term exams and other periodic assessments such as quarterly assessments.

Standardized interim assessments may be appropriate for a variety of uses, including predicting a student’s likelihood of success on a large-scale summative assessment, evaluating a particular educational program or pedagogy, identifying potential gaps in a student’s learning after a limited period of instruction has been completed, or measuring student learning over time.

There are three other types of interim assessments currently in use beyond the “backward looking” interim assessments described above. All are “forward-looking.” One useful but less widely used type is a pre-test given before a unit of instruction to gain information about what students already know in order to adjust plans for instruction before beginning the unit (teachers may do these pre-instruction checks on a more frequent, formative basis). Such forward-looking assessments may be composed of pre-requisite content or the same content as the end-of-unit assessment. A second type of forward-looking assessment is a placement exam used to personalize course-taking according to existing knowledge and skills. Finally, a third type of forward-looking assessment is intended to predict how a student will do on a summative assessment before completing the full unit of instruction. The usefulness of this type of interim assessment is debatable in that it is unlikely to

provide much instructionally relevant information and there is often other information available to determine who is likely to need help succeeding on the end of year summative assessment.

A Note on Classroom Assessment and Accountability

If considerable resources are provided to support classroom-level formative, interim, and summative assessment, there may be a reasonable question as to whether funds are being invested wisely. One temptation may be to hold educators, schools, and/or districts accountable for results on classroom assessments, but such uses are inappropriate for formative and interim assessment, and great care is needed when using classroom summative assessments in such ways. Rather than holding schools and/or teachers accountable for student data gathered from classroom interim and formative assessment, the investment could be evaluated instead by:

- Monitoring the *quality* of formative, interim, and summative classroom assessment practices *rather than outcomes* based on those assessments in such a way that encourages collaboration.
- Requiring teachers and administrators to attend high-quality professional development (PD) on best practices in classroom assessment.
- Monitoring the *degree and quality of administrator support* for teachers to collaborate and improve their formative, interim, and summative classroom assessment practices *rather than outcomes* based on those assessments.

If student *data* from formative or interim classroom assessment are used for educator or school accountability, implementation is likely to be corrupted, and beneficial instructional effects of the investment are likely to be lost.

SECTION 2: DESIRED CHARACTERISTICS AND USES

Introduction

With the background of appropriate characteristics and uses of assessment from Section 1, it is possible to have a coherent presentation of the desired characteristics, uses, and outcomes of assessment as developed by the Task Force.

The Task Force considered that assessment design is always a case of optimization under constraints⁵. In other words, there may be many desirable purposes, uses, and goals for assessment. However, they may be in conflict. Any given assessment can serve only a limited number of purposes well. Finally, assessments always have some type of restrictions (e.g., legislative requirements, time, cost, etc...) that must be weighed in finalizing recommendations.

Task Force members initially were asked to ignore constraints, and identify their desired purposes and goals for assessment and their desired uses of assessment data. Subgroups of Task Force members noted their highest priority uses, and then reviewed the work of other subgroups, asking clarifying questions. After each subgroup's highest priority uses and purposes were reviewed, each individual panelist identified their three highest priorities. The full task force then discussed possible patterns emerging from the activity.

In general, Task Force members desire a Wyoming assessment (system) that is capable of serving the following broad purposes:

- Provide instructionally-useful information to teachers and students (with appropriate grain-size and timely reporting)
- Provide clear and accurate information to parents and students regarding students' achievement of and progress toward key outcomes, such as progress toward meeting grade-level standards and progress toward readiness for post-secondary education and/or career training
- Provide meaningful information to support evaluation and enhancement of curriculum and programs
- Provide information to appropriately support federal and state accountability determinations

Top priority uses and characteristics that were similar were consolidated. In consolidating, important differences in each contributing uses/characteristics were incorporated into the consolidated description. Appendix B provides more detailed information regarding this prioritization activity.

An important outcome of this activity is that no single type of assessment (formative, interim, or summative) is applicable to all of the high-priority desired uses and characteristics. In fact, formative assessment is uniquely able to support two uses/characteristics and summative assessment is uniquely able to support three uses/characteristics. The same is true for level of assessment: classroom-level and state-level assessment are each uniquely able so support three uses/characteristics.

⁵ See Braun (in press).

These outcomes of the Task Force’s work indicate that in order to accomplish the full set of uses and characteristics, **a system of assessments** would be required that span the range of assessment type (formative, interim, and summative) and assessment level (classroom, district, and state). This can be accomplished by combining state and local assessments in a way that they create a coherent system that eliminates unnecessary assessment and provides a consistent picture with complementary characteristics and uses.

A Statewide Summative Assessment or an Assessment System?

As stated above, a single assessment is incapable of meeting the various high-priority characteristics and uses identified by the Task Force. In order to do so, all three types of assessment may be necessary. However, in the same way that a pile of bricks does not make a house, a collection of assessments at the classroom, school, district, and state level is not necessarily a coherent assessment system capable of meeting multiple intended uses⁶.

It is clear that the Task Force desires to respect local control in Wyoming education, maintain the autonomy of individual educators, and provide educators appropriate professional development and ongoing support. Designing a comprehensive assessment system within statutory constraints that also meets the desires listed above is difficult and complex, but not impossible. **Based on these considerable difficulties and complexities, the Task Force was faced with a decision: Recommend a single statewide summative assessment to fulfill statutory requirements or a comprehensive assessment system.**

The Task Force first voted to explore the possibility of a comprehensive assessment system (with a few members expressing reluctance and reserving judgment). After further discussion in later meetings, **Task Force members unanimously voted to make recommendations for a comprehensive assessment system.** As a prelude to the specific recommendations, Task Force members identified issues with the existing state, interim, and district assessments that should be addressed in developing recommendations. They also developed intended outcomes based on those issues. Those issues and intended outcomes are presented in Section 3. A narrative summary of the Task Force recommendations for addressing those issues and achieving the intended outcomes is provided in Section 4. Detailed recommendations to assist in developing one or more requests for proposal (RFP) and in evaluating vendors’ bids on those RFPs are provided in Section 5. Changes to policy necessary to allow for implementation are presented in Section 6.

⁶ See Coladarci (2002).

SECTION 3: INTENDED OUTCOMES

Introduction

In developing recommendations for a new state summative assessment, the Task Force deliberated on issues it intended to address in three areas: state summative assessment, interim assessments, and district assessment systems. The issues identified by the Task Force include the following:

Issues to Be Addressed

Interim Assessment

The Task Force identified incoherence between the existing state assessment and the various interim assessments currently in use as an issue. It is important for the state and interim assessments to provide consistent information about individual students and groups of students to assure that difference seen in the results are not simply artifacts of differences between the tests in terms of format, quality, and content coverage.

State Summative Assessment

Timing and Stability

- The state summative assessment is administered too early in the year to reflect a full year of instruction, and on the flip side results sometimes come too late for use in school improvement activities such as program and curriculum evaluation. The assessment needs to be administered later in the year *and* results need to be returned in time for use in school improvement.
- The use of state test scores for school improvement activities is tenuous because the test or the cut scores on the test change too often. The state assessment needs to remain stable for many years to allow for analysis of policies, programming, and curriculum over time.
- Comparing results from Wyoming state assessment to other states is not possible because the assessment is unique to Wyoming. It is important that Wyoming be able to compare its results with other states with similar content standards to inform state and local policy.

Test Quality

- The quality and usefulness of student achievement and growth reports needs to be improved, given the high-stakes use of state test results. It is important that the state assessment include high-level tasks representative of the kind of teaching we expect from Wyoming educators and learning we expect from Wyoming students.
- It is important for the test to represent both the depth and breadth of the Wyoming state content standards. Multiple-choice-only tests are inadequate in that they signal that Wyoming puts a priority on easy-to-measure knowledge and skills.

Concerns about Appropriate Use

- Educators need adequate professional development in appropriate uses of state assessment data and appropriate preparation for success on the assessment. Teachers need confidence that they can appropriately use state assessment data to improve their own practice.
- Educators need adequate professional development in appropriate uses of state assessment data and appropriate preparation for success on the assessment. Teachers need confidence that they can appropriately use state assessment data to improve their own practice.
- Current use of ACT goes beyond what is appropriate. The ACT is a college entrance examination that is built to measure ACT's college readiness standards. It was not developed to measure the Wyoming state content standards. As such, it is inappropriate to use the ACT as the sole accountability assessment in high school. The use of college entrance assessment scores should be limited to the use for which it has been validated: predicting college success.
- The use of ACT as the sole high school accountability assessment has resulted in confusion about what the high school learning targets are: the official Wyoming state standards or the ACT college readiness standards? Wyoming high school educators need the high school learning targets to be clear in order to appropriately align their instruction to one set of learning targets.

District Assessment Systems

While Wyoming districts have been responsible for developing local assessment systems for a long time, there has been little review of the technical quality of such assessment systems. The Task Force recognized the need for improving the quality of district assessments to increase their usefulness in informing local decisions and for documenting student learning of the basket of goods. The following three general issues were identified:

- Varying levels of coherence of district assessment systems with the state assessment and with interim assessments, leading to confusion in conclusions drawn from the various assessments.
- Varying degrees of quality of district assessment systems.
- Inadequate local capacity to develop and validate high-quality local assessment systems.
- Inadequate evaluation and quality control of local assessment systems.

Intended Outcomes of a Comprehensive Assessment System

Based on desired characteristics and uses of assessment developed in Section 2 and on issues identified above, the Task Force developed intended outcomes of a new Wyoming Comprehensive Assessment System in several broad areas, as shown below.

Integrating Assessment and Instruction

- Prioritize the Wyoming state content standards in a transparent way so that educators clearly know what knowledge and skills will be included on the test and that the complete set of test-eligible content is feasible to teach in the allotted instructional time.

- Improve day-to-day integration of assessment with instruction by encouraging both teacher-level collaboration and material administrative support for initial and ongoing professional development and collaboration at the state, district, and school levels.
- Provide teachers and administrators with timely data on individual students' strengths and weaknesses, and their current and predicted future achievement of desirable outcomes.

Improving Student and Parent Engagement

- Assist students (and their parents) to become more engaged in their own education through a greater knowledge of their strengths and weaknesses and their current (and likely future) achievement of desirable outcomes by providing daily feedback from formative assessment and periodic evaluative data from interim and summative assessment.

Achieving Alignment, Coherence, and Stability

- Achieve alignment of curriculum, instruction, and assessment with the officially adopted Wyoming state standards in every district to ensure that every Wyoming student is provided a high-quality opportunity to learn the “basket of goods.”
- Achieve coherence of local, interim, and state assessments.
- Achieve stability of local and state assessments to allow for a single-minded focus on improving instruction rather than adapting to new assessments.

Improving Student Academic Achievement and Growth

- Better inform educational policy improvement by providing high-quality data, stable across many years, to high-level policymakers.
- Hold schools and districts appropriately accountable for better measured and more desirable student outcomes.
- Provide valid data to local administrators in order to adjust programs and curriculum to target areas of weakness.

Improving the Quality of Assessment

- Improve the quality of district assessment systems.
- Expand beyond multiple choice to include other types of tasks on the state assessment better suited to measuring high-level knowledge and skills.
- Convey to all Wyoming education stakeholders that high-quality writing is a valuable skill that must be effectively taught and learned in Wyoming public schools.

Enhancing the Grade 11 and 12 Experience

- Limit state-required, standards-based, accountability testing to grades 3-10.
- Reserve testing time in grade 11 and 12 for individualized college entrance, work readiness, Advanced Placement (AP), and International Baccalaureate testing.

- Provide freedom in grades 11 and 12 to encourage universal development and use of individualized pathways through a Career & Technical Education (CTE) program and/or college preparation program.
- Provide freedom in grades 11 and 12 for dual enrollment programs strengthen high school ties to community colleges and universities.
- Provide freedom in grades 11 and 12 to smooth students' transitions from high-school to postsecondary education and/or training
- Provide freedom in grades 11 and 12 for students to obtain valuable certificates by the time of graduation.
- Improve equity in options available to all high-school students regardless of location by providing grade 11 and grade 12 options in all Wyoming high schools.

Section 4 provides an overview of the system recommended by the Task Force. Section 5 provides detailed recommendations. Sections 4 and 5 are presented separately because it is difficult to get a coherent picture of what the assessment system would look like from the various detailed recommendations.

SECTION 4: NARRATIVE RECOMMENDATIONS FOR A COMPREHENSIVE ASSESSMENT SYSTEM

Introduction and Context

Wyoming stakeholders have determined that they want an assessment system that will serve multiple purposes, including documenting Wyoming student academic achievement and growth rates as well as supporting local instructional and program evaluation needs. A thoughtfully-designed system of state, local, and classroom assessments will be necessary to achieve these goals. Such a system will yield high-quality data from all levels of the education system to support a variety of purposes. The Task Force strongly supported minimizing redundant assessments while maximizing coherence of the results. The Task Force prioritized the following broad purposes for the Wyoming Assessment System:

- Producing instructionally-useful information for teachers and students,
- Providing clear and accurate information to parents and students regarding students' achievement of and progress toward key outcomes,
- Producing meaningful and useful information for school administrators and policymakers to support evaluation and enhancement of curriculum and programs, and
- Providing appropriate information to support state and federal accountability determinations.

This section of the report describes the Task Force's recommendations for a Comprehensive Wyoming Assessment System, attempting to paint a picture of an assessment system that blends high-quality state and local assessment results to support the multiple purposes described above. Wyoming's educational system, in spite of the centralized funding model, is notably based on local control. Therefore, the Assessment Task Force recommends an approach to assessment that supports the multitude of uses described above, but that strongly values and improves the quality of locally-generated information.

The assessment system recommended by the Task Force is comprised of statewide, standards-based summative assessments in English language arts, mathematics, and science; a set of interim assessments intentionally linked with the summative assessments; district assessments designed to ensure that students have had an opportunity to learn the "basket of goods;" and formative assessment practices controlled at the school and classroom levels. The Task Force supported employing summative assessments that can accurately measure deeper levels of student thinking, but to do so as efficiently as possible so that the summative assessment does not occupy an oversized place in the overall system. The Task Force emphasized that formative assessment is exclusively a local endeavor, but welcomed developing state-district collaborations to support local or regional professional learning opportunities. Finally, the Task Force recognized that the perceived and actual usefulness of any assessment system is limited by the quality of data and reporting capabilities. While the Wyoming Department of Education has made significant strides in capitalizing on modern data visualization techniques to facilitate accurate interpretation of the school accountability results (WAEA), more work is required to develop a reporting structure that enhances the utility of the results from various assessments while minimizing potential misinterpretations.

Proposed Wyoming Assessment System

The Wyoming Assessment Task Force recommends designing and implementing an assessment system that relies on local assessment results to provide rich information to support instructional and evaluative decisions (such as curriculum and program evaluation), while relying on state summative assessments to support accountability decisions. This is done by focusing on improving assessment practice and the quality of data produced by four main assessment system components:

1. **Classroom formative assessment** practices designed and implemented by teachers to inform moment-to-moment monitoring of student learning and allow for immediate adjustment of instruction, and to provide high-quality feedback to engage students in monitoring and furthering their own learning.
2. The **district assessment system** used to document students' opportunities to learn the "basket of goods" can take many forms, ranging from district-selected or -created end-of-course summative to assessments to end-of-unit or similar interim assessments aggregated over the course of a year to produce determinations of student performance in specific courses/grades.
3. **State-supported interim assessments** in state-tested content areas are designed to provide checks on student performance a few times during the school year and/or provide feedback on how well students have learned key clusters of academic knowledge and skills. The Task Force recommends that as part of the contract for the state summative assessment, the state also contract for an interim assessment tied to the summative assessment that local districts may use **as part of district assessment systems**.
4. **State end-of-year or end-of-course summative assessments in grades 3-10** designed to support state school (and perhaps district) accountability decisions, serve program evaluation needs at local, regional, and state levels, and to audit local assessment results.

For these four categories of assessments to work coherently in Wyoming, they must, at a minimum, be designed to measure student learning of the Wyoming content standards in each of the nine required content areas.

Classroom Formative Assessment

The Wyoming Assessment Task Force acknowledged the critical importance of classroom formative assessment practices for improving student learning, but emphatically argued that it should remain relatively silent on recommendations in that area. Task Force members noted that formative assessment is the purview of districts (actually, schools and classrooms) and, for the most part, should not be considered a state program. The Task Force, however, acknowledged that it would make sense for the state and districts (perhaps organized regionally) to partner in providing high-quality professional development to support increasing and improving local formative assessment practices.

District Assessment System

In response to State Supreme Court decisions and legislative mandates, Wyoming requires districts to document that students have had an opportunity to learn the "basket of goods." A comprehensive assessment system must address how the state will monitor student learning of this basket of goods. The combination of district assessment systems and state summative assessments in English language arts, mathematics, and science are required to meet these mandates. The

legislature and State Board of Education have had quality assurance requirements for district assessment systems in place for more than 15 years. In spite of this history, the Task Force members expressed concern about the effectiveness of these requirements and the utility of the feedback and supports provided to districts for improving their assessment systems.

The Task Force noted that district assessments play multiple roles, contingent upon their intended uses. Districts have designed a variety of approaches to meet local needs and work within the constraints of capacity. District summative assessments are expected to be aligned to the relevant Wyoming content standards in the given grade level or course, but the specific assessment approach may vary considerably across districts. For example, districts may choose to use single, large-scale tests at the end of a grade or grade span or they may rely on multiple unit-based assessments tied to the applicable Wyoming content standards. In another example, district assessments may serve both an auditing function for individual teachers' understanding of their students' learning, and a signaling function of the kinds of knowledge and skills that should be prioritized in daily instruction and classroom assessment.

Even so, Task Force members expressed frustration that in spite of the mandate that districts design and implement local assessment systems in at least nine content areas, there was little clarity regarding the state-required purposes and intended uses of these systems. As explained previously, assessments work best when designed for a specific use (in fact, we argue that is the only way that assessments are useful) and if the intended purposes of the district assessment systems are vague, the utility of the results will be limited. Many districts have taken matters into their own hands and have designed assessment systems that meet local needs. This may be appropriate, but it makes it difficult to outline specific quality criteria if the assessments across districts are designed for considerably different purposes. The Task Force strongly recommended having common requirements of assessment quality, but supported local flexibility on specific assessment designs and uses. The Task Force also thought it might be more appropriate to consider having flexibility in design and use become a privilege limited to schools and districts performing well on the school accountability system. On the other hand, the Task Force thought the requirements for district assessments should be tighter when schools within a district have low accountability scores. Further, WDE could require districts with schools receiving low accountability scores to receive training on assessment literacy and learn how to use assessment results to support improvement. In this case, district assessments should be designed to provide more fine grained information than the state assessment.

There was interest among some legislators, as expressed in Enrolled Act 87, in using district or other local assessments for state and/or federal accountability purposes while reducing the amount of statewide summative testing. However, the Task Force declined to move in that direction at this time. Task Force members were concerned that meeting the quality requirements for district assessments to serve accountability uses could overwhelm district personnel. After examining the data and reviewing the existing literature, the Wyoming Assessment Task Force recommends that, at the current time, district assessment results should not be used as part of school accountability determinations. The Task Force acknowledged that such a stance may relegate district assessment results to a lower status than the state assessment. At the same time, Task Force members were concerned that it was not practically feasible in the short term to dramatically improve the quality of district assessments so they could be used as accountability indicators.

However, the Task Force recognized the need for improving the quality of district assessments through the use of multiple strategies including increasing the assessment expertise of those who

reviewed district assessments as part of district accreditation processes and to foster local assessment expertise through state support of district assessment consortia.

Interim Assessments

The Wyoming State Legislature has required and paid for the implementation of a common interim assessment program for all Wyoming school districts. The State supported two administrations of the interim assessment each year—fall and spring—but many districts paid to support winter administration as well. While many district leaders found value in the commercially-selected interim assessment products, using them for a variety of purposes including documenting within-year growth and identifying students in need of remediation, the Task Force members expressed some concern expressed that it was difficult to coherently interpret the results of the interim assessments in light of the summative assessment expectations because the two were designed to measure different learning targets and to do so in different ways (e.g., different item formats).

The Wyoming Assessment Task Force’s major recommendation on the interim assessment was that the State should require the development of an interim assessment system based on the same assessment framework and tied to the same learning targets as the state required summative assessment. Districts may choose to adopt the state-provided interim assessments, and would have local control over how they would administer the tests and use the results. Districts would have the option of purchasing/developing an interim assessment system not tied to the state assessment system, but such districts would be responsible for the costs.

In a critically-important move to help inform WDE’s procurement process the Task Force made additional recommendations regarding the specific interim assessment design. A key consideration for interim assessment design is whether the assessments are “forward-looking,” “backward-looking,” or a “mini summative assessment” design. Forward-looking assessments are provided prior to instruction to gain an understanding of student readiness for learning new concepts and skills. Conversely, backward-looking assessments are those that are designed to help educators and students know how well students learned material that had been taught, generally recently. They can be designed as modules to evaluate student learning of discrete aspects of grade level content (e.g., numbers and operations).

Mini-summative designs are those in which each instance of the interim assessment (2, 3, or 4 or more times each year) is designed to replicate the summative assessment blueprint⁷. Because they are intended to be on the same scale (often a vertical score scale), such mini-summative interim assessment designs are often used for evaluating student growth throughout the year. On the other hand, there is a substantial body of research indicating that vertical scales are not necessary for documenting student progress. Many Task Force members indicated that it is important for interim assessments to “predict” end-of-year summative assessment performance, and thought that the mini-summative designs were the best way to meet this need. However, the technical facilitators (Martineau and Marion) pointed out that it would be relatively easy to create prediction equations for almost any pair of reasonably well correlated assessments.

⁷ A test blueprint is generally in the form of a matrix where the content categories (e.g., standards, objectives) to be tested are represented on one axis and the level of cognitive demand (in the form of process skills or depth of knowledge) required is represented on the other axis. The cells then document the number of test items or score points for each content category by each level of cognitive demand that can be expected to appear on the test.

Task Force members were intrigued by having a set of modules, tied to key aspects of grade-level content, as the potential interim assessment design. In order to keep costs in check, the modules would be focused on a limited number of the major concepts of the discipline (e.g., 3-5 modules) and designed so that districts could administer the modules when and where they fit best within each district's curriculum. The modules offer promise for providing feedback to educators and students on more narrowly-specified sets of knowledge and skills than the broader set of content associated with a mini-summative design. Such modules could also effectively serve an auditing function for district assessments, which should be designed to measure similar knowledge and skills. Finally, a modular approach to interim assessment offers the potential for simultaneously reducing the time associated with the summative assessment and generating more instructionally-useful information for educators. Because this possibility may seem counterintuitive, additional explanation is provided in the footnote at the bottom of this page⁸.

In order to achieve this goal, it may be necessary to customize an existing assessment to some degree. Given the recommendations that follow about not using a custom-designed large-scale summative assessment in Wyoming, existing assessments would need to be capable of a degree of customization without the loss of the benefits that an existing assessment offers. This will likely be

⁸ Subscores serve as achievement reports on subsets of the full set of knowledge and skill represented by a total score. For example, many English language arts summative assessments produce a total score for English language arts, subscores for at least reading and writing, and often finer-grained subscores for topics such as informational and literary reading. Similarly, a mathematics test typically yields an overall math score and potential subscores in topics such as numbers and operations, algebraic reasoning, measurement and geometry, and statistics and probability. One of the greatest challenges in current large-scale summative assessment design is to create tests that are no longer than necessary to produce a very reliable total score (e.g., 5th grade mathematics) while yielding adequately reliable subscores to help educators and others gain more instructionally-relevant information than gleaned from just the total score.

Unfortunately, there is a little known aspect of educational measurement (outside of measurement professionals) that large-scale tests are generally designed to report scores on a "unidimensional" scale. This means that the 5th grade math test, for example, is designed to report overall math performance, but not to tease out differences in performance on things like geometry or algebra because the only questions that survive the statistical review processes are those that relate strongly to the total score of overall math. If the test was designed to include questions that better distinguish among potential subscores, the reliability (consistency) of the total score would be diminished. There are "multidimensional" procedures that can be employed to potentially produce reliable and valid subscores, but these are much more expensive to implement and complicated to ensure the comparability of these subscores and the total score across years. The National Assessment of Educational Progress (NAEP) is the one example of a well-known assessment designed to produce meaningful results at the subscore level, but NAEP has huge samples to work with and more financial resources and psychometric capacity at its disposal than any state assessment. In other words, it is not realistic at this time to consider moving away from a unidimensional framework for Wyoming's next statewide summative assessment, which means that the subscores will unfortunately be much less reliable estimates of the total score than useful content-based reports. This is true for essentially all commercially-available interim assessments as well so that in spite of user reports that they like assessment X or Y because it produces fine-grain subscores useful for instructional planning, any differences in subscores are likely due to error rather than anything educationally meaningful.

In spite of this widely-held knowledge by measurement professionals, every state assessment designer knows that they need to produce scores beyond the total score otherwise stakeholders would complain they are not getting enough from the assessment. Recall that producing very reliable total scores is critical for accountability uses of statewide assessments and, all things being equal, the reliability is related to the number of questions (or score points) on a test⁸. Therefore, most measurement experts recommend having at least 10 score points for each subscore with to achieve at least some minimal level of reliability, so that statewide summative tests tend to get longer to accommodate subscore reporting. Therefore, one way to lessen the time required on the statewide summative assessment is to focus the summative assessment on reporting the total score and use the optional modules for districts that would like more detailed and accurate information about particular aspects of the content domain.

possible by 2018. Another potential benefit that such an approach offers is further reducing the amount of student time devoted to state summative assessments⁹.

The Task Force also discussed types of questions that should appear on the interim assessments. The members knew that using selected-response items (e.g., multiple-choice) to populate the interim assessments would allow for instant reporting and would keep costs down. However, the Task Force recommended that interim assessment questions reflect the types of questions found on the large-scale summative assessment designed to probe students' deep understanding of critical content and skills. At the same time, the Task Force also strongly recommended that the interim assessment scores must be returned to schools within one week of completing the test. This tradeoff would allow for questions that might take a little longer to score than instant multiple-choice items, but might not allow for the full array of extended-response tasks.

Finally, the Task Force issued recommendations around existing and future accountability requirements associated with the interim assessments. The Task Force recommended that requiring districts to implement assessments in order to conduct evaluations of specific programs could easily become unwieldy and result in a hodgepodge of assessments instead of the coherent system that the Task Force is promoting. The Bridges program is a case in point. This intervention program is designed to provide supplemental educational opportunities to traditional educationally-disadvantaged student groups or other students struggling with grade-level knowledge and skills. These opportunities are often provided during the summer, but may be offered after school and on weekends during the regular school year. While well-meaning, the notion of requiring the administration of interim assessments early in the school year to help evaluate the Bridges program has the effect of making the "state" assessment a three times per year event and, most importantly, may miss important aspects of the Bridges program.. It is generally assumed that a fall interim assessment allows for calculation of change in students' scores from spring to fall after experiencing the Bridges summer school. However, as noted above, Bridges funds are commonly used to implement instructional interventions other than summer school, such as weekend programs throughout the school year, meaning that a fall interim test for Bridges evaluation may be limited. It is beyond the scope of this report to discuss alternative evaluation designs for the Bridges program. Rather, the Task Force emphasized that the legislature and other policy bodies should avoid requiring additional assessments without carefully thinking about how such assessments fit within a comprehensive assessment system.

⁹ If districts use modular state-provided interim assessments (see previous footnote) to obtain subscores in each content area, it is not necessary for the state summative assessment to produce anything more than an overall group-level score in each content area for accountability subgroups in each school and district. Subscores provided through modular interim assessments can provide students, parents, and educators with the necessary information to summarize strengths and weaknesses for the purposes of educational decision-making (e.g., planning course-taking, ability grouping, evaluating and enhancing curriculum and programming). Overall group-level scores provided through state summative assessments can provide policymakers with appropriate scores for use in accountability. The reduction in testing time can be achieved by avoiding the need for every student to take every part of the state summative assessment. Rather than every student taking every part of the state summative assessment, each student can be strategically assigned to complete only a portion of the state summative assessment in each content area in such a manner that the entire set of content standards is addressed across each group of students. This allows for the calculation of a group-level outcome for use in accountability rather than requiring the use of complete scores for every individual student.

State Summative Assessment

The Task Force indicated that the state summative assessment must comply with state and federal laws, industry best practices, and professional standards. Further, the assessment should be designed using principled assessment design and to minimize any undue burden on local districts and students. The Task Force strongly recommends that in content areas where it is possible, the state summative assessment selected for Wyoming should be used in at least one other state (preferably many states). There are two reasons for this: to allow for comparison of Wyoming educational outcomes to other states and to encourage a stable state summative assessment over time. In other words, changes to the state summative assessment should be minimized by requiring negotiation with other states and/or a vendor in order to make changes to the assessment system.

The Task Force recommended limiting testing time for state-required summative assessments to no more than *one percent* of the Wyoming required instructional hours for any grade. This translates to a limit of 9, 10.5, and 11 hours of testing time for elementary, middle, and high school grades, respectively. The Task Force did *not* recommend that the full limit of hours be used, only that this should be the maximum allowable. The recommendation is intended to assure that testing time for state summative assessment is kept at a reasonable level and to assure the ability to include questions measuring high-level knowledge and skills on the assessment. State tests are not timed in Wyoming so the Task Force recommended that required testing time be estimated as the amount of time needed for at least 85 percent of students to complete testing. These estimates will improve in accuracy over time.

The Task Force recommended that state, standards-based summative assessments be required in English language arts (including writing) and mathematics in grades 3-10 as well as in science in at least one grades each in elementary, middle s, and high school. These assessments must be designed to fully measure the Wyoming content standards and to assess whether students are on track towards college and career ready outcomes. The Task Force recommends that the grade 10 state summative assessment should count as part of the Hathaway scholarship¹⁰ determinations to explicitly tie the scholarship to the official Wyoming content standards and to assure adequate student motivation.

The Task Force pointed out that it is not appropriate to include all of the Wyoming high school standards on a test given in grade 10, because students still have at least two more years of school remaining. Therefore, the Task Force recommends having the Wyoming Department of Education convene a standards review committee to determine which of the state high school content standards are eligible for testing by the end of 10th grade. Because grades 11 and 12 remain important, the Task Force recommends that district assessment systems be required to cover the Wyoming high school content standards that do not appear on the state summative assessment. The Task Force noted that such prioritization could occur easily with a custom assessment program, but would have to be negotiated if the state procures a consortium, collaborative, or other existing assessment system.

¹⁰ The Hathaway scholarship is a program where Wyoming high school students who complete a required set of courses, have a certain grade point average (GPA), and achieve the required composite score on the ACT. There are various levels of the scholarship award ranging from \$1640 to \$840 per semester (for 2015 graduates) depending on the specific GPA and ACT scores. It was beyond the scope of the Task Force's work to recommend exactly how the grade 10 scores may be included as part of the Hathaway determination, but the Task Force was confident that this was not an insurmountable problem.

The Task Force also recommends that the state continue to fund in-school administration of a college entrance examination in grade 11. However, the Task Force argued that career readiness was as important as or more important than college readiness in many parts of Wyoming. Therefore, the Task Force recommended requiring all students to participate in *either* a college entrance examination or an analogous career readiness assessment. The provision of an in-school opportunity for college entrance or career readiness testing (rather than a traditional Saturday administration) is intended to maximize the number of students thinking about post-secondary opportunities.

The recommendations to have the last required state standards-based summative assessment at the end of 10th grade is designed to encourage students to specialize during their last two years of high school. The lack of state mandated standards-based testing in grade 11 and 12 is designed to help junior and senior students focus on highly individualized pathways through either college preparation (e.g., through Advanced Placement [AP], dual enrollment, or other programs) or specific career/technical areas where students may become “concentrators.” It also facilitates the transition from high school into college or career training by strengthening the connection between grades 11-12 and post-secondary education or training.

In order to improve reporting timelines for use in school improvement and other evaluation activities, the Task Force recommends administering state summative assessments online except in isolated situations with emergent needs for paper and pencil. Safeguards for assuring a successful transition to online testing are described near the end of this section of the report. The Task Force recommends administering the summative tests in a three-week window near, but not at, the end of the school year to maximize the amount of instructional time before the test, but also to assure return of results in time to support summer school improvement activities and district program evaluation needs.

The Task Force recommends that the state summative assessments serve both an auditing function for district assessment results and a signaling function of the kinds of knowledge and skill that should be prioritized in district assessments (e.g., deeper levels of thinking).

However, the task force is concerned that including too many performance or other extended-response tasks on the state summative assessment may lead to unacceptable testing times. Therefore, the Task Force strongly recommends that the state summative assessment include the minimum number of questions necessary to both signal the types of assessment tasks the state would like to see on classroom and district assessments and ensure that the state assessments can provide information about student learning of the full depth of the content standards.

Supports and Conditions

To improve fidelity of implementation at the classroom, school, district, and state levels, the Task Force noted that certain supports are critical.

Data and Reporting Systems

The Task Force recommends the use of a comprehensive assessment system to maximize the coherence of information produced from various assessment tools. However, without a well-designed and implemented reporting system, the hopes for a comprehensive assessment system will

fall well short. The world of data visualization has opened up exciting new possibilities for placing useable information in the hands of users in ways they can easily understand. Score reports are the only ways assessment designers communicate with stakeholders, yet it is often the last thing attended to in design deliberations¹¹. Therefore, the Task Force strongly recommends that Wyoming devote the resources necessary to produce a high-quality digital reporting system that capitalizes on modern data visualization techniques and facilitates accurate assessment interpretations while minimizing opportunities for misconceptions. Such a reporting system could be included in vendors' bid in response to the state assessment RFP, but the Task Force is aware that such systems would likely come from more specialized vendors. The Task Force commended WDE's efforts in reporting the results of Wyoming Accountability in Education Accountability system (WAEA), but wanted to go much further to help users understand the assessment results and potential educational implications of the scores.

The Task Force recognized that sophisticated reporting techniques are still limited by the quality and grain size of the information provided by state assessments. The state assessment results are necessarily based on a broad survey of the standards and not detailed content information suitable for guiding instructional actions. Therefore, an ideal reporting system would integrate state assessment and accountability results, interim assessment scores, and local (district and classroom) information into a coherent picture of student learning. It would also have the capacity to house student work samples for understanding student learning over time in terms of the content and quality of their work. There are obvious ownership (state/district), privacy, and capacity issues to work out with stakeholders to assure comfort and the effective use of the system.

Assessment Literacy

Having high-quality and intuitively useable reporting systems is a big step toward improving assessment literacy. Unfortunately, it is probably not enough. The Task Force recognized WDE's current efforts to promote formative assessment practices, but recommended expanding the state's efforts to promote assessment literacy and effective assessment. It is beyond the scope of this report to fully outline approaches to meet these goals. The Task Force recommends implementing a thoughtful approach or set of approaches to improve local assessment practices and products (e.g., classroom and district assessments).

Evaluation

Finally, the state should contract for an ongoing evaluation of (1) the quality of the state assessment; (2) the degree to which intended outcomes are being achieved; (3) the degree to which anticipated and unintended consequences have been observed and minimized (for the unintended, negative consequences); and (4) after three to five years, a summary report including potential improvements to the system to address any issues identified.

Ensuring a Successful Transition

The Task Force recommends a multi-year transition strategy to ensure a successful transition to online state summative assessment and high-quality interim assessment systems. **Allowing the full three years from the time of acting upon these recommendations is critical to assuring that**

¹¹ Attributed to Ron Hambleton.

the transition is successful. The first all-online administration of the state summative assessment will take place in the spring of 2018 and the transition must be smooth. The Task Force recommends a comprehensive set of safeguards to assure a smooth transition, as follows:

1. Schools and districts will be notified as soon as possible that they must be ready for online assessments in the spring of 2018.
2. As soon as possible, the state will contract for a high-quality comprehensive technology infrastructure audit for the state as a whole and for every school and district. The state audit will at a minimum cover adequacy of the state internet backbone. District audits will at a minimum cover adequacy of available bandwidth, stability of connections to the state backbone and/or other networks. School audits will at a minimum cover adequacy of available bandwidth, stability of connections to district/state systems, adequacy of wireless school network capacity, adequacy of the number of devices capable of administering the assessment, and the adequacy of the operating systems used on those devices.
3. The state contractor will work with each school district to assist in performing the audit (including fully conducting the audit if necessary) to assure a consistent application across all districts.
4. The state contractor will produce reports for the state, district, and school. The report will identify specific gaps in technology infrastructure and minimum actions that must be taken to close them.
5. All appropriate state agencies that will support school technology infrastructure should pledge their support for preparing all schools for online assessment by spring 2018 and clearly describe what forms their support will take.
6. At least ten months in advance of the first online administration, all schools, districts, and the state contractor will conduct a simultaneous load test simulating all of Wyoming's students logging on and taking the test simultaneously to attempt to "break" the system. Before the first administration, any breaks or near breaks in the system as a result of the load test will be used to increase capacity in any areas necessary.
7. A paper and pencil option must be available to address isolated emergent needs that cannot be resolved in a reasonable amount of time to allow for online testing.
8. Schools should have reasonable flexibility on scheduling testing within the test window to accommodate the use of online assessment in case the schools possess a limited number of devices.
9. Students should be provided with adequate experience in the classroom using the same or very similar devices as those that will be used for the tests. At a minimum, this should include specific focus on navigating a screen and keyboarding. The Department of Education should gather a workgroup of educators to develop guidelines for providing adequate experience.

SECTION 5: DETAILED RECOMMENDATIONS

Introduction

The Task Force urges that the recommendations included in this report generally not be written into Wyoming statute or into Wyoming Department of Education rule. They further urge that any existing statute or rule contradictory to recommendations in this report be eliminated or appropriately modified to allow for full implementation of these recommendations.

Rather than writing the recommendations in this report into statute or rule, the Task Force urges that these recommendations instead be embodied in a Request for Proposals (RFP) to be issued so that vendors can bid on providing the services required to implement the system. This understanding is important in that it allow for minor adjustments. However, it would be reasonable to require general compliance with these recommendations where it is feasible to do so and where an unanticipated compelling reason to choose a different course does not arise.

Classroom Formative Assessment

The Wyoming Assessment Task Force acknowledged the critical importance of classroom formative assessment practices for improving student learning, but emphatically argued that other than briefly discussing formative assessment in this report, the Task Force should remain relatively silent on the issue. Task Force members noted formative assessment is the purview of districts (actually, schools and classrooms) and for the most part should not be part of the “state” comprehensive assessment system. The Task Force, however, acknowledged that it would make sense for the state and districts (perhaps organized regionally) to partner in providing high-quality professional development to support high-quality local formative assessment practices.

District Assessment System

As the major issues identified with district assessment systems are uneven quality and uneven coherence with state assessment, several recommendations address quality control and information flow:

- To facilitate information flow between districts and the state, a two-way data exchange should be implemented. Flowing from the state to the district, state-level data are transmitted to local district electronic systems, where teachers and administrators can access individual and aggregate state, local, and classroom data for their students. Flowing from district systems to the state are district-level standards-based designations from district summative assessments. These links can also be used to audit district-level standards-based designations and identify districts with local assessment systems that may need improvement. The Department of Education will need to work with stakeholders to develop protocols for data exchange and security to ensure student privacy and the appropriate use of local data for audits.
- District data systems should be developed to house samples of students’ work along with scores for each of the required standards and skills to document learning of the “basket of goods.”

- The state should contract with a vendor with experience in high-quality interim and summative assessments including performance tasks and projects to measure high-level knowledge and skills. This vendor should fill two roles: (1) provide district and school personnel with statewide professional development in developing high-quality interim and summative assessments, and (2) for districts that request assistance in developing or refining local systems, provide that assistance on a cost optional basis.
- To improve quality and assure consistency of reviews, the state should contract with one or more qualified professionals to perform audits of district assessment systems as a part of the accreditation process.
- The state should incentivize and/or support collaborative efforts among districts to improve the quality of locally-developed assessment tasks and the quality of data use for informing educational decisions. This could be modeled after the WY BOE Assessment Activities Consortium. This could include hosting for educators to obtain access to intact assessments, banks of high-quality tasks and test questions, and appropriate professional development on using the resources.

Because considerable improvements in district assessment systems would be required to support high-stakes use, the workgroup recommends NOT using the district assessment results as an indicator in WAEA at this time.

Interim Assessment

Governing Principles

The Task Force recommends that the state support an interim assessment system to encourage consistency across the state. The use of interim assessments should be governed by the following principles:

- To reduce required testing time, districts should not be required to administer any interim assessments, but may choose to integrate interim assessments into its district assessment system if integration is appropriate¹².
- Districts choosing to integrate the state-provided interim assessment into their district assessment systems would not be responsible for the cost of the assessment. Districts choosing to administer a different interim assessment would do so at their own expense.
- The interim assessment supported by the state should be coherently tied to the state summative assessment. It should be constructed to measure the same content standards,

¹² Requiring districts to implement assessments in order to conduct evaluations of specific programs could easily become unwieldy and result in a hodgepodge of assessments instead of the coherent system that the Task Force is promoting. The Bridges program is a case in point. While well-meaning, the notion of requiring the use of interim assessments administered early in the school year to evaluate the Bridges program has the effect of making the “state” assessment a twice per year event and, most importantly, may miss important aspects of the Bridges program. It is generally assumed that a fall interim assessment allows for calculation of change in students’ scores from spring to fall after experiencing the Bridges summer school. However Task Force members reported that Bridges funds are commonly used to implement instructional interventions other than summer school such as weekend programs throughout the school year, meaning that the usefulness of a fall interim test for Bridges evaluation is minimal. It is beyond the scope of this report to discuss alternative evaluation designs for the Bridges program. Rather, the Task Force emphasized that the legislature and other policy bodies should avoid requiring additional assessments without carefully thinking about how such assessments fit within a comprehensive assessment system

and should use the same types and formats of test questions to assure a consistent experience for students and educators across state summative and interim assessment.

- To achieve competitive pricing and coherence, the interim assessment should be procured as part of the summative assessment RFP process.
- To provide an outside audit of the district assessment results, interim assessments should provide a check on the big ideas associated with the grade level learning targets.

Two “Flavors” of Interim Assessment

Because each district assessment system is uniquely designed to meet local needs, the Task Force recommended that the state-provided interim assessment should be made available in two forms to allow for maximum flexibility.

- A “mini-summative” version in which the interim assessment is a shorter version of the end-of-year state summative assessment (e.g., the interim assessment blueprint is representative of the summative assessment blueprint, but results in a shorter test¹³). This allows for monitoring students’ growth within a school year on an overall content area and for predicting student performance on the end-of-year summative test.
- A module-based version in which the blueprint of the summative assessment is broken into 3-5 subsets of content categories, and each interim assessment module measures only one subset. Each module should allow for at least two subscores to be reported within the subset. This allows for measuring achievement of mid-sized units of instruction.

Flexibility

To meet varying needs in the different district assessment systems, the Task Force recommended considerable flexibility in the timing and use of interim assessments as deemed appropriate by districts, schools, and/or teachers, including, but not limited to:

- Module-based unit pre-test for planning and differentiating instruction.
- Module-based unit post-test for measuring achievement of module content¹⁴.
- Module-based unit post-test for identifying remedial needs.
- Module-based unit test for predicting achievement on the end-of-year summative test.
- Module-based unit interim assessment for measuring student growth on module content.
- Mini-summative on prior-grade content for a new student without prior test scores.
- Mini-summative for predicting achievement on the end-of-year summative test.
- Mini-summative for measuring student growth on the full content area.

Item and Task Types

¹³ A test blueprint is generally in the form of a matrix where the content categories (e.g., standards, objectives) to be tested are represented on one axis and the level of cognitive demand (in the form of process skills or depth of knowledge) required is represented on the other axis. The cells then document the number of test items or score points for each combination of content category and level of cognitive demand that can be expected to appear on the test.

¹⁴ This use could reduce educator workload in creating end-of-unit, mid-term, and or final exams for grading purposes.

The Task Force recognized the importance of the interim assessment mirroring the summative assessment as much as possible to assure that complex knowledge and skills are measured on both. The Task Force also recognized that near-immediate reporting is needed to maximize the usefulness of interim assessments. The inclusion of complex item types (see **Alignment to the Wyoming State Standards** on page 33) means that human scoring may be required, which increases the time between completing an assessment and reporting. To address this conflict, the Task Force recommends the following compromise:

- Interim assessment results should be returned no more than one week after completion of an assessment.
- All items types used on the summative assessment should also be included in the interim assessment, so long as they do not preclude returning interim assessment results in no more than one week.

State Summative Assessment

Governing Principles

Quality is of critical importance if assessments are to be appropriately used to inform educational decisions. To assure that Wyoming is able to procure a high-quality assessment, the Task Force recommends the following:

- To avoid loss of negotiating power and to improve the ability to meet Wyoming's needs, assessment products should not be named in statute, rule, or policy. Nor should statute, rule, or policy so tightly define requirements that only one product is qualified.
- The technical quality of the assessment should be well-documented according to research and/or best practices as referenced by some or all of the following:
 - Principled assessment design (e.g., *Evidence Centered Design*¹⁵, *Knowing What Students Know*¹⁶)
 - Universal Design for Learning¹⁷
 - The AERA/APA/NCME *Standards*¹⁸
 - CCSSO/ATP *Best Practices* for Statewide Assessment¹⁹
 - Applicable state and federal law and regulation
 - Federal peer review requirements

Standards-Based Assessment vs. College/Career Entrance Assessment

To address confusion about the official Wyoming state standards in high school, to maintain the benefits of a college entrance examination, and to provide greater freedom for Juniors and Seniors to pursue individualized pathways, the Task Force recommends that a distinction be made between assessments up to grade 10 and assessment after grade 10 as follows:

¹⁵ Mislavy & Riconscente (2006).

¹⁶ Pellegrino, Chudowsky, & Glaser (2001).

¹⁷ Thompson, Johnstone, & Thurlow (2002).

¹⁸ APA, AERA, & NCME (2014).

¹⁹ CCSSO & ATP (2013).

- Assessment after Grade 10.
 - Reserve grade 11 and 12 for college entrance, work skills, Advanced Placement, and International Baccalaureate assessment. Do not add standards-based state summative assessment in grade 11 or 12.
 - To provide schools incentives to help upper level high school students develop highly individualized pathways through a career and technical education program or a college preparation program, do not use grade 11 and 12 assessments for school accountability purposes.
 - Require grade 11 students to take *either* a college entrance examination or a work skills examination. This should be administered in school on a regular school day.
 - The Department of Education should be provided with funding for a contract to provide students with detailed information about their career/college interests and development of individualized high-school pathways.
- Assessment in Grades 3-10
 - Require standards-based, state summative accountability assessment in grades 3-10.
 - The Department of Education should be provided with funding for a contract to conduct studies to develop predictive relationships between the grade 9 and 10 assessments and the college readiness and work skills assessments.
 - To ensure both (1) student motivation on the grade 10 test, and (2) alignment of the Hathaway scholarship criteria with the official Wyoming content standards, include the grade 10 assessment in the criteria for Hathaway scholarship eligibility, with opportunities to retest in grades 11 and 12²⁰.

Testing Time

In combination with eliminating the requirement to use a state-provided interim assessment, the Task Force recommends limiting the amount of time that may be required for state summative assessment.

- Limit actual testing time for state-required summative assessment to no more than 1% of the required instructional hours for a given grade level (based on Chapter 22 of Wyoming Department of Education rules, this is a maximum of 9, 10.5, and 11 hours of testing time for elementary, middle, and high school, respectively)²¹.
- “Actual testing time” means the time that students are actually responding to assessment tasks (not additional time used for test preparation, breaks, gathering students, logging students, or reading test instructions)²². Because Wyoming state assessments are not timed, “actual testing time” should be based on estimated testing time needed for 85% of students to complete the test. These estimates should be updated annually based on actual test administration.

²⁰ There are several ways in which this may be accomplished. The Task Force was particularly interested in an approach in which students may demonstrate qualification on the grade 10 assessment, the college entrance assessment, or the work skills assessment. Whether such an approach is appropriate will need to be determined once concrete plans for high school assessment have been put in place through a competitive bidding process.

²¹ Required testing time may be less than these limits.

²² This definition of “actual testing time” is provided to avoid district-to-district variation in the time devoted to activities wrapped around actual testing.

Test Timing and Test Windows

In order to balance maximizing the amount of instructional time before state summative assessments and typical end-of-year school activities, and the need to receive results in time for school improvement activities, the Task Force recommends the following:

- State testing should occur during a three- to four-week testing window which is the same for every grade, with the last allowable testing day being in the first half of May.
- All aggregate reports (other than statewide aggregate reports) should be available by August 1 to facilitate school improvement activities (with consideration that in the first year of any new program, reports are likely to be delayed).
- Acting within the constraints of the first bullet in this list, the Department of Education should survey districts to set a first allowable and last allowable testing day for each year. If possible, start and end times should be later to maximize instruction before assessment, but should also consider typical year-end school activities and the time needed to return data to schools in time for use in school improvement activities.
- Acting within the constraints of the first bullet in this list, the Department of Education should work with a committee of stakeholders to finalize testing windows (e.g., the first and last allowable testing days each year) and to address local needs for flexibility in scheduling assessment activities²³. If possible, start and end dates should be later to maximize instruction before assessment, but should also consider typical year-end school activities and the time needed to return data to schools in time for use in school improvement activities. This committee of stakeholders should include school and district staff with two sets of responsibilities: (1) calendaring, and (2) managing state assessment activities..

Content Coverage

To ensure compliance with federal laws and to signal the importance of the core content areas of Language Arts (including Writing), Mathematics, and Science, the Task Force recommends the following:

- Require assessment of Language Arts and Mathematics in every grade.
- Require coverage of Writing (as a part of Language Arts) in *at least* one grade each in the elementary, middle, and high school grade spans. If it is possible to do so within the limits for testing time, include writing in each of grades 3-10.
- Require coverage of Science in *at least* one grade each in the elementary, middle, and high-school grade spans.
- To clearly identify what content is eligible to appear on the grade 10 test in each content area, the Department of Education should facilitate a standards review committee with the charge of specifying which of the Wyoming content standards are expected to be taught and learned by end of grade 10. The committee should be composed of K-12 content specialists, district curriculum directors, and higher education content specialists. Any remaining high-school content should be covered in district assessment systems.

²³ For example, allow for flexibility in length of test sessions to coincide with the length of class periods (to avoid unnecessary disruption of daily instructional activities).

Alignment to the Wyoming State Standards

In order to eliminate confusion about the official Wyoming state standards in high school, and to signal the importance of complex knowledge and skills described in the Wyoming state standards, the Task Force recommends the following:

- The grade 3-10 assessments should be aligned to the depth and breadth of Wyoming’s state content standards, including complex knowledge and skills that are not easily measured.
- The assessment should include both multiple choice items and more complex item types better suited to measuring more complex knowledge and skills (e.g., enhanced multiple choice, technology enhanced items, short constructed response, extended constructed response, performance tasks). However, the number of more complex item types included in the assessment must allow for meeting the testing time limits.
- To avoid market restriction, vendors proposing “naked” writing tasks should not receive lower scores in determining whether they are qualified bidders. However, after qualified bidders have been identified, vendors proposing writing tasks that are embedded in a text-based response should receive extra credit over vendors proposing naked writing tasks²⁴.

Moving Assessment Online

The Task Force recommends that test administration be moved fully online to expedite return of assessment results and the use of data in school improvement activities (such as evaluation and adjustment of instructional approaches, curriculum, and programming). However, given Wyoming’s problematic history with online assessment, the transition *must* be smooth. Several safeguards will be put in place to assure a smooth transition, listed below. The most important of these is that the new assessment system should be developed and implemented over three years. If these recommendations are acted upon quickly, a new assessment system could be in place by spring of 2018. The recommended safeguards to assure a smooth transition to fully-online assessment include the following:

- Schools and districts will be notified immediately that they must be ready for online assessment in spring of 2018.
- The state will immediately contract for a high-quality comprehensive technology infrastructure audit for the state as a whole and for every school and district. The state audit will at a minimum cover adequacy of the state internet backbone. District audits will at a minimum cover adequacy of available bandwidth, stability of connections to the state backbone and/or other networks. School audits will at a minimum cover adequacy of available bandwidth, stability of connections to district/state systems, adequacy of wireless school network capacity, adequacy of the number of devices capable of administering the assessment, and the adequacy of the operating systems used on those devices.

²⁴ This assumes a bid evaluation process in which vendor bids are first scored to determine whether they meet a threshold for qualifying to provide the state with assessment services, followed by a review of the qualifying bids for a few areas in which select vendors may receive extra credit for proposing “value added” beyond the requirements of the request for proposals (RFP).

- The state contractor will work with each school district to assist in performing the audit (including fully conducting the audit if necessary) to assure a consistent application across all districts.
- The state contractor will produce a public report including sections for the state as a whole (including a summary of district and school reports), each district (including a summary of each school report), and each school. The report will identify specific gaps in technology infrastructure in each section of the report and identify minimum actions that must be taken to close those gaps.
- After the full set of audit reports has been produced, it may be necessary for the legislature to consider whether there are any critical, targeted funding needs to fill the identified gaps.
- All appropriate state agencies that will support school technology infrastructure should pledge their support for preparing all schools for online assessment by spring 2018 and clearly describe what forms their support will take.
- At least ten months in advance of the first online administration, all schools, districts, and the state contractor will conduct a simultaneous load test simulating all of Wyoming's students logging on and taking the test simultaneously to attempt to "break" the system. Any breaks or near breaks in the system as a result of the load test will be used to increase capacity in any areas necessary before the first administration.
- A paper and pencil option must be available to address isolated emergent needs that cannot be resolved in a reasonable amount of time to allow for online testing.
- Schools should have reasonable flexibility on scheduling testing within the test window to accommodate the use of online assessment with a limited number of devices (e.g., the length and number of test sessions for each student).
- It will be communicated often to both parents and educators that prior to taking assessments online, students should be provided with adequate experience in the classroom using devices they will take the test on. This should include at a minimum specific focus on navigating a screen and keyboarding. The Department of Education should gather a workgroup of educators to develop guidelines for providing adequate experience.

Claims that Must Be Supported for Individual Students

In order to support important educational decisions made by parents, students, and teachers, the Task Force recommends that the assessment must support the following claims for each individual student:

- How each student achieves relative to Wyoming standards, including more difficult to measure, high-level knowledge and skills.
- How each student achieves in producing high-quality writing (requires at least two extended written responses per student to support this claim).
- How each student gained in learning relative to peers.
- That student achievement and growth scores are accurate across the range of student achievement, meaning that:
 - Scores are generally free of floor or ceiling effects.
 - Scores support claims about whether novice, typical, and advanced students are being well educated.

Claims that Must Be Supported for Classrooms, Schools, Districts, and the State

In order to support important educational decisions made by teachers, administrators, policymakers, and the public, the Task Force recommends that the assessment must support the following claims for each classroom²⁵, school, district, and the state:

- The magnitude of achievement and growth gaps for key demographic groups (e.g., sex, race/ethnicity, economic disadvantage, special education, and English learners).
- The change in achievement and growth gaps over time.
- The percentage of Wyoming students meeting proficiency targets.
- The percentage of Wyoming students meeting growth targets adequate to remain proficient (for already proficient students) or to achieve proficiency (for not yet proficient students) within a reasonable number of years.
- Produces valid and reliable group reports (at the class, school, district, and state level) on strengths and weakness in both proficiency and growth in a small number of sub-areas of each content area. This supports school improvement activities, post hoc evaluation of instructional practices, curriculum, and programming, and high level policies. This could be accomplished using green/yellow/red light reports that show for each group the sub-areas in which a group's achievement is better than, similar to, or worse than its overall content area achievement²⁶.

Reporting

Without thoughtfully designed and useful reports, the quality of the assessment system is moot. To assure that investment in the quality of the assessment is returned, the Task Force recommends the following:

- Reports must be designed to meet the needs of the following four groups of stakeholders with similar interests:
 1. Students and parents
 2. Teachers
 3. School and district leadership teams
 4. Business community, media, State School Board, State Superintendent, Joint Legislative Education Committee, Legislature at large, Governor, and general public
- Individual student reports must be designed with stakeholder groups 1 and 2 in mind.
- Aggregate reports (e.g., classroom and school reports) showing individual student data must be designed with stakeholder groups 2 and 3 in mind.
- Aggregate report showing group summary data must be designed with all four groups of stakeholders in mind.

²⁵ Access to classroom-level aggregate reports should be limited to educators responsible for that classroom to protect student privacy.

²⁶ For example, group average subscores can be compared to overall scores within a content area to identify whether in each sub-area, the group perform better than, similar to, or worse than they did in the overall content area. Each of those group average scores can also be compared to the thresholds for the different performance levels.

- Unless it is possible to adequately serve the needs of multiple stakeholder groups with a single report format, each report should be developed with a format specific to each audience.
- The format and elements of each report should be determined by conducting focus groups and/or multiple rounds of workshopping, with a focus on the following for each report element:
 - Identifying the critical “so-what” message(s) for the intended audience(s).
 - Assuring that the “so-what” message(s) are clearly and transparently conveyed.
 - Designing reports to minimize probable misinterpretations.
 - Assuring consistency with AERA/APA/NCME standards for score reporting²⁷.
- The reporting system should allow for teachers to receive dynamic individual reports for just their current students, and aggregate reports for their current and past students.
- The reporting system should allow for each audience to obtain the desired information using intuitive navigation and assistance in finding reports to answer specific questions. Report users should be able to retrieve data to answer their questions with a minimum number of clicks through guided selection of options. Where access to data is appropriate, report users should be able to easily retrieve data about achievement and growth for individual students and demographic groups at the student, classroom, school, district, and state level; with simple navigation between levels.

Avoiding an Exclusive Wyoming Assessment

In order to provide stability, cost savings, enhanced quality, and comparability of Wyoming test results to other states, the Task Force recommends the following:

- Each content area test must be used in some form in at least one other state (preferably several other states) for the following reasons:
 - Provide stability by requiring changes to the assessment to be negotiated with at least one other state and/or vendor.
 - Facilitate comparison of results from the Wyoming assessment to results from other states.
 - Reduce cost through multi-state collaboration.
 - Improve technical quality through the increased capacity and expertise in a multi-state collaboration.
- To maximize market competition, the ability to meet Wyoming’s needs, and negotiating power, recommendations in this section should be required only where there are at least two options available.

Wyoming Educator Participation in Ongoing Development

In order to improve the fit of the assessment to the Wyoming context, and to assure understanding of the assessment by Wyoming educators, the Task Force recommends the following:

- To avoid market restriction, vendors whose proposals are not consistent with recommendations in this section should not receive lower scores in determining whether

²⁷ APA, AERA, & NCME (2014).

they are qualified bidders. However, after qualified bidders have been identified, vendors whose proposals are consistent with recommendations in this section should receive extra credit²⁸.

- Although avoiding an exclusive Wyoming assessment means that development will already be completed, it is desirable that Wyoming educators have the opportunity to be involved in ongoing development and maintenance of the assessment.
- Wyoming educators have substantive say in ongoing development activities including item development, item review, rangefinding, and other development activities.
- Wyoming educators have the opportunity to review test questions for specific Wyoming sensitivities.
- If there are alternative test questions available to replace those flagged as problematic by Wyoming educators, WDE is able to replace the flagged questions.
- Wyoming educators are involved in scoring student responses requiring human scoring for tests completed by Wyoming students
- The Wyoming Department of Education defines and oversees Wyoming educator involvement.

Test Security

In order to avoid the considerable stress and disruption to students, educators, and families caused by test security breaches, the Task Force recommends the following:

- The Department of Education must develop a high quality policy document and associated training using industry standards on test security.
- The policy document and training must include clear policies, protocols, and guidelines to comprehensively address test security in all aspects of testing including at least the following areas:
 - Professional development
 - Prevention of test security breaches
 - Detection of test security breaches (including balancing protection for whistleblowers and minimizing the impact of malicious allegations)
 - Investigating potential security breaches
 - Protocols for evaluating evidence to make conclusions
 - Protocols for appeals of conclusions
 - Follow-up activities to a substantiated or suspected security breach
- The Department of Education’s test administration vendor must assist with test security to supplement agency capacity in each of the areas listed in the previous recommendation.
- The Department of Education’s test administration vendor must document its own security procedures throughout its processes.

Data Security and Privacy

²⁸ This assumes a bidding process in which vendor bids are first scored to determine whether they meet a threshold for qualifying to provide the state with assessment services, followed by a review of the bids for a few areas in which select vendors may receive extra credit for proposing “value added” beyond the requirements of the request for proposals (RFP).

In order to protect the privacy of individual student data and to comply with Federal student privacy law, the Task Force recommends that the vendor must document that its corporate policies on data security and privacy comply with all applicable state and federal statutes and regulations, that those policies are adequately strong to prevent data security breaches, and that those policies are rigorously enforced.

Program Evaluation

In order to determine whether the State's investment in a new comprehensive assessment system is achieving the intended results, the Task Force recommends the following:

- The state should contract for an independent summary report evaluating the degree to which the intended outcomes of the state summative assessment have been realized after five years of implementation.
- The evaluation should include the following at a minimum:
 - The quality of the state assessment
 - The degree to which intended short-, mid-, and long-term outcomes are being realized
 - The degree to which anticipated unanticipated unintended consequences have been observed
 - Should this be an ongoing evaluation, or does this invite instability?
- To monitor for concerns before and after the five-year evaluation, and to make recommendations as needed, the Department of Education should empanel from this point forward a statewide assessment policy advisory committee (PAC) that meets at least twice a year. This panel should include teachers, administrators, technology coordinators, and assessment coordinators. Because stability of the state assessment is paramount, the first activity of this committee should be defining thresholds for making changes. These definitions should strongly privilege stability of the system over time, meaning that thresholds concerns about the assessment must meet before changes are made must be high.

Specialty Assessments

The Task Force focused its efforts on designing a coherent assessment system for the general student population in the content areas comprising the basket of goods. The Task Force also recognizes the importance of coherence of its recommendations in four additional specialty areas:

- Alternate assessment for students with significant cognitive disabilities
- English proficiency assessment for English language learners
- Early literacy assessment in grades K-3
- YCTA career and technical education concentrator assessments

However, the Task Force was largely composed of general educators, and recognized the need for specialists in each of these areas to make appropriate recommendations for these specialty assessments. Therefore, the Task Force recommends that in each of these three areas, the Department of Education convene small committees of experts to review the recommendations for state summative assessment presented in this report. Those committees should then make recommendations for those assessments to be coherent with the general content area assessments by determining which of the recommendations in this report are appropriate for those assessments,

which are inappropriate, which need to be modified, and to identify any additional recommendations that may be needed.



SECTION 6: POTENTIAL QUALIFYING PRODUCTS

The Task Force put a premium on ensuring assessment quality, practical usefulness of assessment data, and on state-provided assessments not being exclusive to Wyoming. At the same time, the Task Force and the State Board of Education at its September 23, 2015 meeting expressed concern about whether the recommendations in this report may unreasonably reduce the number of potential qualified bidders. **While the Task Force presents these companies as potential bidders, this in no way means that the company would either respond to a Wyoming RFP or that they would be able to meet the requirements of the RFP.** Any potential Wyoming assessment vendor would have to provide evidence that their product can meet the requirements outlined in the RFP.

Language Arts and Mathematics

Table 6.1 below presents the potential companies and products would be likely or possibly available for Language Arts and Mathematics. This information is based on the knowledge of the two authors as a result of their work in other states and knowledge of the industry.

Table 6.1. *Likely and possibly qualifying products.*

Source	Type of Source	Status as of Spring 2015
ACT Aspire	Test Vendor	Administered in 2015 in two (2) states
Data Recognition Corporation	Test Vendor	Ready for use
Educational Testing Service	Test Vendor	Under development
Measured Progress	Test Vendor	Under development
PARCC	Consortium of States	Administered in 2015 in eleven (11) states
Smarter Balanced	Consortium of States	Administered in 2015 in eighteen (18) states
University of Kansas	State University	Administered in 2015 in two (2) states
Utah	State sells test items	Administered in 2015 in two (4) states

Based on Table 6.1, it appears that there are sufficient sources of likely and possibly qualifying products to assure that there is adequate and competitive bidding. We list in red some potential sources in Table 6.1 even though (1) no documentation is currently available for the products they have developed or are in the process of developing, and (2) no other state is currently using products from those sources for statewide summative assessment. We include these potential sources because by the time a request for proposals (RFP) is issued, these vendors may have adequate documentation and their products may have been adopted by at least one other state.

Finally, for Language Arts and Mathematics there are a few additional important considerations about collaboration with each potential source that may be probed in an RFP and in scoring bids on the RFP. Wyoming must consider the degree of control it wants in any new assessment system. Several of the potential products—such as ACT Aspire, DRC, ETS, Measured Progress, University of Kansas, and Utah—would afford Wyoming very little, if any, control over the assessment program. On the other hand, if Wyoming became a governing member of an assessment consortium (PARCC or Smarter Balanced), it may have a limited amount of influence over the nature of the assessment system. In either case, Wyoming may extend its influence by convincing

other states of the importance of its position and together with other states recommend a change to the assessment program.

Second, the division of labor differs across potential assessment providers. In the case of ACT Aspire, DRC, ETS, Measured Progress, PARCC, and University of Kansas, the assessment provider is solely responsible for product development and for test administration, scoring, and reporting; and the state is responsible for overseeing contract performance. Smarter Balanced is responsible for product development and monitoring consistency across member states and states are responsible for procuring a state-specific vendor for test administration, scoring, and reporting and for monitoring the contract performance of that vendor. , On the other hand, PARCC manages all assessment activities centrally. States such as Florida, Tennessee, and Arizona have purchased the rights to use Utah test items in 2015, but there is no cross-state collaboration beyond that financial transaction.

Science

Science is addressed separately because whereas there is considerable similarity of the Wyoming state standards in Language Arts and Mathematics to those of many other states, the Wyoming state standards in Science are unique. Therefore, there may or may not be sources with qualified products (meaning that an exclusive Wyoming science assessment may be needed). The potential assessment options available for science will depend on the new science content standards adopted by the Wyoming State Board of Education.

Of the sources listed in Table 6.1, ACT Aspire, Utah, and the University of Kansas offer science assessments. The DRC, ETS, and Measured Progress products may include science assessments when they become available. PARCC and Smarter Balanced products do not include science assessments. The degree to which the ACT Aspire, Utah, and University of Kansas science assessments are aligned to the Wyoming state science standards is unknown. The degree of alignment of existing science assessments would need to be independently evaluated to determine whether collaboration provides a benefit over keeping an exclusive Wyoming science assessment.

Another potential avenue for collaboration is that one or more other states' standards in Science *may* be adequately similar to the current Wyoming standards to make collaboration worthwhile. The degree of alignment between Wyoming and other states' science standards would need to be evaluated, as would the alignment of other states' science tests to the Wyoming state science standards to determine whether collaboration is worthwhile. There is at least one collaborative effort, organized by the Council of Chief State School Officers (CCSSO) underway currently now to support assessment of the Next Generation Science Standards (NGSS). This is currently focused on the development of an item bank so the state would need to hire a vendor to develop and administer the rest of the assessment.

Finally, it may be wise to wait to update the Wyoming science assessment until new science standards have been adopted in Wyoming for two reasons. First, investment in a new science assessment may be poorly spent if changes to Wyoming state science standards are considerable. Second, once new science standards have been adopted, they can be compared to those adopted by other states to identify states with sufficiently similar science standards to make collaboration across states more desirable. Finally, depending on the instructional shifts required by any new standards, the state may choose to adjust the timing of the new assessment to best accommodate the required instructional shifts.

SECTION 7: RECOMMENDATIONS FOR POLICY COHERENCE

The Task Force took great care in ensuring that the recommendations put forth in this report are technical and practically sound. However, the Task Force is aware and concerned that several of the recommendations contradict existing statute. Prior to pointing out specific statutes that will need to be amended or repealed in order to implement the recommendations issued here, we offer general guidelines for legislating assessment requirements.

The Task Force spent considerable time discussing and trying to outline a coherent and efficient assessment system for Wyoming. One of the key features of a coherent assessment system is that each assessment in the system is designed to measure the same learning targets in complimentary ways. Further, in order to create an efficient system that minimizes redundancy, each assessment must be carefully designed to produce the intended inferences and to thoughtfully occupy a place in the overall system. It is easy to start adding assessments to meet specific needs (e.g. to support the evaluation of the Bridges program), but this can quickly lead to an incoherent and inefficient set of assessments that no longer function as a system.

Therefore, the Task Force strongly recommends that the legislature create statutes to set broad goals and articulate the intended uses of assessments (e.g., measuring student growth, for use in school accountability determinations). The legislature should prioritize creating a coherent, comprehensive, and efficient assessment system designed to measure student learning of Wyoming content standards and to support school improvement efforts. On the other hand, the legislature should avoid legislation regarding the specifics of assessment design (e.g., types of items to be included on the assessment) or even requiring assessments for specific purposes (e.g., requiring a 3rd grade reading assessment). The Task Force is aware that each time the legislature adds an assessment (e.g., ACT) or adds a specific requirement (e.g., multiple-choice items only), it is for well-intentioned reasons often in response to constituent concerns. Unfortunately, while every action might be well-intentioned, when we look back after a few years, a once coherent assessment is no longer so.

Designing and implementing a stable, efficient, and coherent assessment system requires high levels of technical and practical knowledge. Therefore, we compliment the legislature for appointing the Assessment Task Force, a representative group of citizens, to try to bring more coherence and stability to the Wyoming assessment system. Further, statute tends to last longer than rules and they are often much more difficult to change, especially considering that the Wyoming legislature is in session only 20 or 30 days each year, while the State Board of Education meets monthly to allow for more rapid modification of rules and requirements.

With that framework, we outline the following recommended changes to existing statute to allow the recommendations presented here to be enacted.

1. *W.S. 21-2-202 (a)²⁹: administering a standardized, curriculum based, achievement college entrance examination, computer-adaptive college placement assessment and a job skills assessment test selected by the state superintendent to all students in the eleventh and twelfth grades throughout the state in accordance with this paragraph.* This clause basically requires the ACT and a placement exam such as Accuplacer. The Task Force recommendations would still require the provision of a college entrance or work readiness exam, but the Task Force made no such recommendation for a

²⁹ Also found in W.S. 21-3-110

placement exam. Such an exam may be useful once students enroll in a postsecondary institution, but not as part of the state assessment system. Further, the language of “curriculum based, achievement college entrance exam” is a nod to ACT’s marketing as curriculum measure with SAT as an “aptitude” test. This is simply not true. ACT is no closer to Wyoming’s standards than the SAT. Therefore, the Task Force recommends a more neutral requirement for a college entrance and career readiness exam.

2. *W.S. 21-2-304 (iv)*³⁰. *Effective school year 2013-2014, and each school year thereafter, require district administration of common benchmark adaptive assessments statewide in reading and mathematics for grades one (1) through eight (8) in accordance with W.S. 21-3-110(a)(xxiv).* The Task Force recommended the optional (at the district level) use of interim assessments, but most importantly to have the interim assessment procured as part of the state assessment RFP. The Task Force did not recommend the use of an adaptive assessment, per se, but for an interim system that best fit the instructional needs of districts. This is an example of what might be considered over-specification of the interim assessment requirement.
3. *W.S. 21-2-304 (v) (B). Effective school year 2012-2013, and each school year thereafter, be administered in specified grades aligned to the student content and performance standards, specifically assessing student performance in reading and mathematics at grades three (3) through eight (8). In addition, the statewide assessment system shall assess student performance in science in grades four (4) and eight (8).* As seen earlier in this report, the Task Force is recommending administering the state assessment system in English language arts and mathematics continuously in grades 3-10. The Task Force suggests leaving the science assessment in place until new content standards are adopted.
4. *W.S. 21-2-304 (v) (C). In addition to subparagraph (a)(v)(B) of this section, measure student performance in Wyoming on a comparative basis with student performance nationally.* While this requirement has not been implemented previously, except through the National Assessment of Educational Progress (NAEP), the Task Force supports the intention of this clause.
5. *W.S. 21-2-304 (v) (E). Use only multiple choice items to ensure alignment to the statewide content and performance standards.* The legislature already knows this is a problematic clause, but has been waiting for recommendations from the Task Force to deal with this clause. The Task Force has made clear that it wants to be able to include the types of test questions necessary to fully and deeply measure the Wyoming content standards and not be limited in the types of questions available to use. This is also an example of the type of specification that should not be in statute.
6. *W.S. 21-3-401: Reading assessment and intervention.* The Task Force did not have the time or the specific expertise necessary to address the reading assessment requirements, but recommends that WDE convene an expert advisory panel to make recommendations regarding K-3 reading assessment. While there is often a desire to produce comparable (standardized) data, early childhood reading assessments must yield information so that teachers can understand students’ unique strengths and weaknesses. This might require the use of individually-administered assessments tied to each district’s specific reading program.

³⁰ Also found in W.S. 21-3-110

7. *W.S. 21-13-334 (b)(iv) Implement a structured common assessment evaluation of program effectiveness.* While not specified in this clause, the common, adaptive interim assessment required under *W.S. 21-2-304 (iv)* has been the defacto common assessment used as the evaluation instrument for this program. As noted in this report, the Task Force argued that the timing of the common interim assessment was not necessarily appropriate for providing data to evaluate the efficacy of the program. Therefore, the Task Force recommends removing this requirement and replacing it with a requirement for districts to provide an appropriate evaluation of their specific program. WDE should be charged with providing guidance to districts on how best to collect evaluation data tied to the specific requirements of each program.

There are likely other statutes related to statewide and district assessment requirements, but the statutes outlined above are the highest priority targets for modification in order to implement the Task Force recommendations.

REFERENCES

- AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Braun, H. (Ed.). (in press). *Meeting the Challenges to Measurement in an Era of Accountability*. Washington, DC: National Council on Measurement in Education.
- CCSSO & ATP. (2013). *Operational Best Practices for Statewide Large-Scale Assessment Programs*. Washington, DC: Authors.
- Coladarci, T. (2002). Is it a house...or a pile of bricks? Important features of a local assessment system. *Phi Delta Kappan*, 83(10), 772-774.
- Michigan Department of Education. (2013). *Report on Options for Assessments Aligned with the Common Core State Standards*. Retrieved June 20, 2015, from http://www.michigan.gov/documents/mde/Common_Core_Assessment_Option_Report_441322_7.pdf.
- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-Centered Assessment Design. In T. M. Haladyna, & S. M. Downing (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC. Retrieved September 4, 2015, from http://www.nap.edu/openbook.php?record_id=10019&page=R1.
- Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief*. Retrieved June 20, 2015, from http://www.nciea.org/publication_PDFs/PolicyBriefFINAL.pdf.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments (Synthesis Report 44)*. Minneapolis, MI: University of Minnesota, National Center on Educational Outcomes. Retrieved September 5, 2015, from <http://www.cehd.umn.edu/NCEO/onlinepubs/synthesis44.html>.
- Wiley, E. C. (2008). *Formative Assessment: Examples of Practice*. Retrieved August 11, 2015, from http://ccsso.org/documents/2008/formative_assessment_examples_2008.pdf.

APPENDIX A: UNDERSTANDING FORMATIVE ASSESSMENT

Definition of Formative Assessment

Formative assessment has also been called formative instruction. The purpose of formative assessment is to evaluate student understanding against key learning targets, provide targeted feedback to students, and adjust instruction on a moment-to-moment basis.

In 2006, the Council of Chief State School Officers (CCSSO) and experts on formative assessment developed a widely cited definition (Wiley, 2008):

Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievements of intended instructional outcomes (p. 3).

In addition, Wiley (paraphrased from p. 3) lists five critical attributes of formative assessment:

1. They are based on clear articulations of learning goals as steps toward an ultimate desirable outcome.
2. Learning goals and the criteria for success are clearly identified and communicated to students in language they can understand.
3. Students are frequently provided with feedback directly linked to the learning goals and criteria for success.
4. Students engage in self- and peer-assessment against the criteria for success.
5. Students and teachers jointly own (collaborate on) monitoring student progress over time.

While the practice of formative assessment in general embodies these five attributes, not every example of formative assessment incorporates every attribute. The definition and five critical attributes are based on research linking such practices to student learning gains. The core of the formative assessment process is that it takes place during instruction (i.e., “in the moment”) and under full control of the teacher to support student learning while it is developing. Thus, formative assessment is an integral part of instruction; instruction need not be paused to engage in formative assessment. This embedded assessment is done through diagnosing on a very frequent basis where students are in their progress toward fine-grained learning targets such as those covered by a single class period. This ongoing diagnosis shows both teachers and students where gaps in knowledge and skill exist, and helps both teacher and student understand how to close those gaps.

The definition and critical attributes make clear that formative assessment is not a product, but a process tailored to the details of ongoing instruction to individual students. Effective formative assessment practices occur very frequently, covering very small units of instruction (such as part of a class period). If tasks are presented, they may vary for students depending on where they are in their learning. However, formative assessment processes often occur during regular and targeted questioning of students in small or large groups, observing students as they work in groups and/or engage in tasks. Formative assessment practices may be facilitated using certain technology and related tools. There is a strong view among some scholars that because formative assessment is tailored to the specific context of the classroom and to individual students that results cannot be

meaningfully aggregated or compared. Many of these scholars question whether the observations from formative assessment should even be scored.

Another implication is the critical importance of providing frequent feedback to individual students. Providing each student such frequent and targeted feedback develops his or her ability to continuously monitor the quality of their own work against a clear learning target. It is this targeted and frequent feedback to students that is the most crucial part of the formative assessment process³¹.

The nature of formative assessment implies that the frequently used term *common formative assessment* is a result of confusion about the nature of formative assessment. Other types of assessment may be used formatively for periodic progress monitoring (e.g., to inform mid-course corrections or modifications to curriculum and programming), but only formative assessment as described above is capable of informing instruction on a moment-to-moment basis. Effective formative assessment is tailored to a specific instructional plan and a specific group of students at defined points in their attainment of learning targets. The critical characteristics of formative assessment practices should be common across all teachers, and tools teachers use to implement formative assessment may be common across many teachers, but formative assessment is too tailored to a unique classroom to be common.

Data gathered through formative assessment have limited to no use for evaluation or accountability purposes such as student grades, educator accountability, school/district accountability, or even public reporting that could allow for inappropriate comparisons. There are at least four reasons for this: (1) if carried out appropriately, the data gathered from one unit to the next, one teacher to the next, one moment to the next, and one student to the next will not be comparable; (2) students will be unlikely to participate as fully, openly, and honestly in the process if they know they are being evaluated by their teachers or peers on the basis of their responses; (3) for the same reasons, educators will be unlikely to participate as fully, openly, and honestly in the process; and (4) the nature of the formative assessment process is likely to shift in such a way that it can no longer optimally inform instruction.

These implications create a distinct difference from summative and interim assessment (described below), which are intended to assess student achievement after an extended period of learning. Simply giving students an assessment in the classroom does not mean that the assessment is formative. Use of assessment evidence in a formative manner requires teachers to achieve insight into individual student learning in relation to learning targets, to provide effective feedback to students about those insights, and to make instructional decisions based on those insights. During the formative assessment process, feedback to students and student involvement is essential. Teachers seek ways to involve the student in “thinking about their thinking” (metacognition) to use learning evidence to close the gap and get closer to the intended learning target.

Because there is a great deal of confusion over what constitutes formative assessment, the next part of this appendix provides vignettes of formative assessment in practice. The four vignettes describe the work of four different educators to help readers to better understand what is meant by “formative assessment.”

³¹ See Sadler (1989).

Vignettes of Formative Assessment in Practice³²

High School – Chemistry Mid-Period Check In

As part of instructional planning, a high school chemistry teacher develops both true and false statements related to a micro-unit covering a half hour in high school chemistry. Statements were strategically developed to assess whether students hold anticipated misconceptions. Following the micro-unit, students show thumbs up, thumbs down, or thumbs to the side to indicate whether each statement is true, false, or they don't know. Based on the prevalence of thumbs down and to the side, the teacher may select one of at least four options:

1. Reteach that micro-unit using a different instructional plan the next day.
2. Use pre-planned strategies to address a small number of misconceptions.
3. Strategically group students who put thumbs down or to the side with confident students to discuss their conclusions and monitor group discussions.
4. Work briefly with a one or two students needing additional assistance while the rest of the class engages in the next activity.

Middle School – English End of Period Check In

At the beginning of a seventh grade English class period, a middle school English teacher shares with her students what the three learning targets are for the day. At the end of the period, she asks each student to fill out and hand in a slip confidentially rating their attainment of each learning target in one of the following four categories:

1. I can teach this.
2. I can do this on my own.
3. I need some help with this.
4. I don't get this at all.

The teacher adjusts the next day's lesson plan by creating a simple task asking small groups of students to practice a learning target on which about half the students felt confident. The small groups are strategically selected to include students that are both confident and not confident with the learning target. She also reviews with the entire class another learning target on which few students felt confident. To do so, she asks two students to explain their approach on a specific problem. After gauging current understanding, she decides whether to instruct on that learning target again using a different strategy and different examples than the previous day.

Elementary School – Monitoring Development of Mathematical Understanding

After a successful unit on simple two-digit addition (without regrouping), an elementary school teacher wants students to learn both a regrouping algorithm and why the algorithm works. He demonstrates to his students that their current knowledge and skills are inadequate to accurately deal with two-digit addition requiring regrouping. He does this by assigning small groups of students to solve a problem either using the addition algorithm they already know or by using counting objects. In a subsequent whole-class discussion, the teacher highlights the conflicting answers and asks his

³² Informed by Wiley (2008).

students to think about how place value place might explain why the groups got different answers. He then asks each small group to work on developing its own solution to the problem. After visiting and probing each group to survey current understanding and developing strategies, he asks strategically chosen groups to share their developing solutions, and builds post-activity instruction on the regrouping algorithm around them.

High School – English Capstone Project

As a capstone project for a unit on persuasive writing, a high-school English teacher assigns her students to individually write a persuasive essay incorporating each of the unit learning targets. Each student is to:

- Choose a position on a controversial topic important to him,
- Identify reliable resources for information on his position and a contrary position commonly taken on the topic,
- Summarize the arguments for both positions,
- Use the logical devices taught in the unit to argue for his position,
- Use logical tools to argue the logical superior of his position, and
- Incorporate work in all five previous steps into a coherent persuasive essay.

The teacher divides the capstone project into four subunits (with associated assignments):

1. Choosing a topic, a personal position, an opposing position, and identifying reliable resources;
2. Summarizing arguments for at least two positions on the topic;
3. Arguing for the personal position and against an opposing position on a logical basis;
4. Incorporating into a complete and coherent persuasive essay.

Along with other formative practices, the teacher spends class time making each sub-unit's learning targets explicit and instructing on them. She also uses class time on the day each assignment is due to have students peer-review each other's work, focusing on the learning targets and working on revisions. As assignments are turned in, the teacher provides formative feedback based on the learning target rather than grading each assignment. Only after providing at least one round of formative feedback on each assignment does the teacher grade the final product. She does this to ensure that the formative feedback fulfills its purpose and her evaluation of each student's performance represents what was learned by the end of the unit.

APPENDIX B: ONE-PAGE SUMMARY OF FORMATIVE, INTERIM, AND SUMMATIVE ASSESSMENT

	Formative Assessment	Interim Assessment	Summative Assessment
Characteristics	<ul style="list-style-type: none"> Facilitate effective instruction (does not pause instruction) Learning goals and criteria are clear to students Students self-/peer-monitor progress toward learning goals Students and teachers receive frequent feedback Jointly controlled by each teacher and her students Covers a micro unit of instruction Very frequent (e.g., multiple times per period) Tailored to a set of students and an instructional plan Might be comparable for a classroom, but not beyond <i>Not a product (e.g., quiz, test, bank of questions/ tests)</i> 	<ul style="list-style-type: none"> Pauses instruction for evaluation Controlled solely by a teacher, school, district, or state (or by a consortium of teachers, schools...) Covers a mid-sized unit of instruction Somewhat frequent (e.g., weekly to quarterly) Administered before and/or after a mid-sized unit Based on who controls assessment, results may be comparable across students, teachers, schools, districts, and/or states A product 	<ul style="list-style-type: none"> Pauses instruction for evaluation Controlled solely by a teacher, school, district, or state (or by a consortium of teachers, schools...) Covers a macro unit of instruction (e.g., semester, course, credit, grade) Infrequent (e.g., yearly, finals week) Administered after completing a macro unit Based on who controls assessment ,results may be comparable across students,..., and/or states A product
Uses	<ul style="list-style-type: none"> Engage students in learning/metacognition through frequent feedback and self-/peer-evaluation Monitor moment-to-moment student learning Diagnose individual students' immediate instructional needs Diagnose immediate group instructional needs Immediately adjust instruction Differentiate instruction Self-evaluate micro-unit instructional effectiveness <i>Student results from formative assessment are not appropriate for use in grading or accountability; however, ratings of the quality of formative assessment practice may be appropriate for use in accountability</i> 	<ul style="list-style-type: none"> Evaluate achievement after a mid-sized unit Monitor progress within a macro-unit (e.g., semester, course, credit, grade) Corroborate formative assessment Pre-test to tailor unit instructional plans for the group and individual students Identify post-unit remedial needs Mid-course self-evaluation and adjustment of teacher classroom practices Mid-course evaluation and adjustment of school and district policies and programs Predict performance on summative assessment Grading (and possibly accountability) 	<ul style="list-style-type: none"> Evaluate achievement after a macro unit Monitor progress across multiple macro-units Corroborate interim assessment Evaluate readiness for the next macro unit After-the-fact evaluation/adjustment of broad instructional practices by individual teachers and of curriculum/programming policies by administrators Predict later student outcomes Grading and accountability
Examples	<ul style="list-style-type: none"> Following a micro-unit, students show thumbs up/thumbs down to indicate whether statements developed around anticipated misconceptions are true. Based on prevalence of misconceptions, the teacher reteaches parts of his lesson using a different instructional strategy, strategically groups students to discuss their conclusions, or works briefly with one or two students. At the end of class, students hand in a slip confidentially rating their attainment of each learning target as: (1) <i>I can teach this</i>, (2) <i>I can do this on my own</i>, (3) <i>I need some help with this</i>, or (4) <i>I don't get this at all</i>. The teacher adjusts her next-day group assignments and planned activities accordingly. 	<ul style="list-style-type: none"> Classroom unit quizzes and homework Individual and group unit projects Pre-unit exams of unit pre-requisites Pre-unit exams of unit content End of unit exams Mid-term exams Marking period exams not covering a full macro-unit Quarterly assessments District placement tests 	<ul style="list-style-type: none"> Classroom final exams, projects, and papers School or district final exams, projects, or papers District/state assessments for testing out of a credit District graduation/diploma-endorsement tests Typical state accountability tests High school equivalency tests District graduation tests College admission tests

APPENDIX C: DETAILED HIGHEST PRIORITY USES AND CHARACTERISTICS

The Task Force’s highest priority uses and characteristics are presented in detail in Table B1 below. These uses and characteristics were evaluated by the facilitators using the definitions and appropriate uses of formative, interim, and summative assessments discussed in Section 2 of this report. The evaluation also incorporates differences between classroom-, district-, and state-owned assessments to show the complexity of an assessment system that would be needed to fulfill all of the Task Force’s highest priority uses and characteristics. This evaluation is reflected in additional elements added to Table B1. Those elements identify whether each type and level of assessment has full, some, minimal, or no applicability to the use or characteristic in each row. In addition, in each row the applicability of the various types and levels of assessment to each use or characteristic is briefly explained.

Table B1. *Task Force Highest Priority Uses and Characteristics.*

Total ¹ Score	Number of Votes by Priority			Desired Uses and Characteristics of Wyoming Assessment	Applicability ²					
					Type			Level		
	1 st	2 nd	3 rd		Formative	Interim	Summative	Classroom	District	State
38	10	3	2	Provide information to parents, students, and educators regarding individual student achievement and growth within and across years, including readiness for the next level in a student's K-12 progression - Classroom formative: continuous achievement/growth/readiness data on micro-units - Classroom/district/state interim: periodic achievement/growth/readiness data on mid-sized units - Classroom/district/state summative: yearly achievement/growth/readiness data on macro-units	●	●	●	●	●	●
27	6	4	1	Provide feedback on progress toward standards to inform instruction on more than a yearly basis - Classroom formative: continuous achievement and progress data inform daily instruction - Classroom/district/state interim: periodic unit achievement & progress data informs remediation - District/state summative: interim results might be rolled up for summative determinations	●	●	●	●	●	●
16	0	5	6	Allow for comparisons within the state and across states - State interim: provides within-state comparability if adopted statewide - State summative: provides within-state comparability - State interim/summative: provides cross-state comparability if a multi-state assessment is used	○	●	●	○	○	●
13	2	2	3	Provide reliable and valid data to evaluate program/curriculum effectiveness and alignment to standards - District/state interim: can provide information to inform within- and between-year evaluations - District/state summative: can provide information to inform between-year evaluations	○	●	●	○	●	●

Total ¹ Score	Number of Votes by Priority			Desired Uses and Characteristics of Wyoming Assessment	Applicability ²					
					Type			Level		
					Formative	Interim	Summative	Classroom	District	State
11	3	1	0	Be student-centered (e.g., student is not a number) - Classroom formative: micro-unit diagnostic data to tailor instruction - Classroom/district/state interim: unit diagnostic data to tailor remediation - Classroom/district/state summative: macro-unit data to inform critical yearly decisions	●	●	○	●	●	●
8	0	3	2	Encourage collaboration and sharing best practices - Classroom formative/interim/summative: foster teacher collaboration on teacher practices - District/state interim/summative: foster teacher collaboration on using non-classroom data - District/state interim/summative: foster educator collaboration on curriculum/programming - Limit use of classroom assessment for evaluation to quality of practices and support for collaboration	●	●	●	●	●	●
7	1	2	0	Continually inform instruction with timely feedback - Classroom formative: continual micro-unit diagnostic data to inform daily instruction - Classroom/district/state interim: periodic unit data to inform post-unit remediation	●	○	○	●	○	○
6	1	1	1	Validly inform decisions about post-secondary education/training - State summative: likely to provide based on ties to post-secondary outcomes (onerous for a district)	○	○	●	○	○	●
2	0	0	2	Consistency over time to facilitate the intended outcomes of assessment in Wyoming - District interim/summative: stable longitudinal data can improve decision making - State interim: stable longitudinal data can improve decision making - State summative: likely to improve decision-making because of school/district accountability uses	○	●	●	○	●	●
X				Number of desired uses/characteristics with unique and full applicability	2	0	3	3	0	3
				Number of desired uses/characteristics with full applicability	4	3	5	4	2	5
				Number of desired uses/characteristics with some applicability	1	4	1	1	4	3
				Number of desired uses/characteristics with unlikely applicability	0	1	2	0	2	1
				Number of desired uses/characteristics with no applicability	4	1	1	4	1	0

- Each panelist identified one characteristic as her highest priority, second highest priority, or third highest priority. These were given scores of 3, 2, and 1 respectively. The scores were summed across panelists to give a total score for each desired use/characteristic.
- , ○, ○, and ○ indicate desired uses or characteristics for which the type or level of assessment has full applicability, some applicability, minimal or unlikely applicability, and no applicability, respectively.

APPENDIX D: MINI-SUMMATIVE VS. MODULAR INTERIM ASSESSMENT DESIGNS

As an aid in understanding assessment design, we first describe the general hierarchical format that content standards take by providing an example from grade-5 mathematics:

Content Category
<p>Operations & Algebraic Thinking</p> <ul style="list-style-type: none"> Write and interpret numerical expressions <ul style="list-style-type: none"> <i>Use parentheses, brackets, or braces...</i> <i>Write simple expressions that record calculations...</i> Analyze patterns and relationships <ul style="list-style-type: none"> <i>Generate...numerical patterns...given rules...</i>
<p>Number & Operations in Base Ten</p> <ul style="list-style-type: none"> Understand the place value system <ul style="list-style-type: none"> <i>Recognize [digit values increase tenfold when one place... left]</i> <i>Explain patterns in... when multiplying by powers of 10...</i> <i>Read, write, and compare decimals to thousandths</i> <i>Use place value understanding to round decimals to any place</i> Perform operations...to hundredths <ul style="list-style-type: none"> <i>Fluently multiply multi-digit whole numbers...</i> <i>Find whole-number quotients of whole numbers...</i> <i>Add, subtract, multiply, and divide decimals to hundredths...</i>
<p>Number & Operations—Fractions</p> <ul style="list-style-type: none"> Use equivalent fractions...to add and subtract fractions <ul style="list-style-type: none"> <i>Add and subtract fractions with unlike denominators...</i> <i>Solve [fraction word problems by comparison...]</i> Apply and extend...multiplication and division <ul style="list-style-type: none"> <i>Interpret a fraction [as a division problem]...</i> <i>[Extend whole number] multiplication to...fractions...</i> <i>Interpret multiplication as scaling (resizing)...</i> <i>Solve...problems [with] multiplication of fractions...</i> <i>[Extend division to involve unit fractions]</i>
<p>Measurement & Data</p> <ul style="list-style-type: none"> Convert like measurement units [in the same] system <ul style="list-style-type: none"> <i>Convert among different sized measurement units...</i> Represent and interpret data <ul style="list-style-type: none"> <i>Make a line plot to display [data with fractional units]...</i> Geometric measurement: understand...volume <ul style="list-style-type: none"> <i>Understand volume as an attribute of solid figures...</i> <i>Measure volumes by counting unit cubes...</i> <i>Relate volume to [multiplication and division]...</i>
<p>Geometry</p> <ul style="list-style-type: none"> Graph points on the coordinate plane to solve... <ul style="list-style-type: none"> <i>Use [two] perpendicular lines...to define a coordinate...</i> <i>Represent... points in the first quadrant...</i> Classify two-dimensional figures...on...properties <ul style="list-style-type: none"> <i>[Know category] attributes [apply] to all sub-categories...</i> <i>Classify...figures in a hierarchy based on properties</i>

To aid in explanation, the broadest content categories (at the top of the hierarchy) are displayed in bold. Sub-categories are indented presented in the same color as the broad category they belong to. Sub-sub-categories are further indented and presented in italics.

In a simplified version of test design, the number of test questions or score points that come from each sub-sub-category is clearly specified to reflect the relative importance of each category. For example, if every sub-sub-category were considered equally important, a reasonable test design might specify that every sub-sub-category be measured using two test questions, resulting in the following hypothetical summative test design:

Content Category	# of Items
Operations & Algebraic Thinking	6
Write and interpret numerical expressions <i>Use parentheses, brackets, or braces...</i>	4 2
<i>Write simple expressions that record calculations...</i>	2
Analyze patterns and relationships <i>Generate...numerical patterns...given rules...</i>	2 2
Number & Operations in Base Ten	14
Understand the place value system <i>Recognize [digit values increase tenfold when one place... left]</i>	8 2
<i>Explain patterns in... when multiplying by powers of 10...</i>	2
<i>Read, write, and compare decimals to thousandths</i>	2
<i>Use place value understanding to round decimals to any place</i>	2
Perform operations...to hundredths	6
<i>Fluently multiply multi-digit whole numbers...</i>	2
<i>Find whole-number quotients of whole numbers...</i>	2
<i>Add, subtract, multiply, and divide decimals to hundredths...</i>	2
Number & Operations—Fractions	14
Use equivalent fractions...to add and subtract fractions <i>Add and subtract fractions with unlike denominators...</i>	4 2
<i>Solve [fraction word problems by comparison...]</i>	2
Apply and extend...multiplication and division	10
<i>Interpret a fraction [as a division problem]...</i>	2
<i>[Extend whole number] multiplication to...fractions...</i>	2
<i>Interpret multiplication as scaling (resizing)...</i>	2
<i>Solve...problems [with] multiplication of fractions...</i>	2
<i>[Extend division to involve unit fractions]</i>	2
Measurement & Data	10
Convert like measurement units [in the same] system <i>Convert among different sized measurement units...</i>	2 2
Represent and interpret data <i>Make a line plot to display [data with fractional units]...</i>	2 2
Geometric measurement: understand...volume <i>Understand volume as an attribute of solid figures...</i>	6 2
<i>Measure volumes by counting unit cubes...</i>	2
<i>Relate volume to [multiplication and division]...</i>	2
Geometry	8
Graph points on the coordinate plane to solve... <i>Use [two] perpendicular lines...to define a coordinate...</i>	4 2
<i>Represent... points in the first quadrant...</i>	2
Classify two-dimensional figures...on...properties <i>[Know category] attributes [apply] to all sub-categories...</i>	4 2
<i>Classify...figures in a hierarchy based on properties</i>	2
Total	52

A *mini-summative interim assessment design* is intended to reasonably replicate the summative assessment experience with the exception of being shorter. For example, on an interim assessment with five testing opportunities, this could be accomplished by measuring each content standard with 1 rather

than 2 items, giving the following mini-summative interim assessment design, making each interim assessment half as long as the summative assessment:

Content Category	# of Items on Interim Assessment				
	1	2	3	4	5
Operations & Algebraic Thinking	3	3	3	3	3
Write and interpret numerical expressions	2	2	2	2	2
<i>Use parentheses, brackets, or braces...</i>	1	1	1	1	1
<i>Write simple expressions that record calculations...</i>	1	1	1	1	1
Analyze patterns and relationships	1	1	1	1	1
<i>Generate...numerical patterns...given rules...</i>	1	1	1	1	1
Number & Operations in Base Ten	7	7	7	7	7
Understand the place value system	4	4	4	4	4
<i>Recognize [digit values increase tenfold when one place... left]</i>	1	1	1	1	1
<i>Explain patterns in... when multiplying by powers of 10...</i>	1	1	1	1	1
<i>Read, write, and compare decimals to thousandths</i>	1	1	1	1	1
<i>Use place value understanding to round decimals to any place</i>	1	1	1	1	1
Perform operations...to hundredths	3	3	3	3	3
<i>Fluently multiply multi-digit whole numbers...</i>	1	1	1	1	1
<i>Find whole-number quotients of whole numbers...</i>	1	1	1	1	1
<i>Add, subtract, multiply, and divide decimals to hundredths...</i>	1	1	1	1	1
Number & Operations—Fractions	7	7	7	7	7
Use equivalent fractions...to add and subtract fractions	2	2	2	2	2
<i>Add and subtract fractions with unlike denominators...</i>	1	1	1	1	1
<i>Solve [fraction word problems by comparison...]</i>	1	1	1	1	1
Apply and extend...multiplication and division	5	5	5	5	5
<i>Interpret a fraction [as a division problem]...</i>	1	1	1	1	1
<i>[Extend whole number] multiplication to...fractions...</i>	1	1	1	1	1
<i>Interpret multiplication as scaling (resizing)...</i>	1	1	1	1	1
<i>Solve...problems [with] multiplication of fractions...</i>	1	1	1	1	1
<i>[Extend division to involve unit fractions]</i>	1	1	1	1	1
Measurement & Data	5	5	5	5	5
Convert like measurement units [in the same] system	1	1	1	1	1
<i>Convert among different sized measurement units...</i>	1	1	1	1	1
Represent and interpret data	1	1	1	1	1
<i>Make a line plot to display [data with fractional units]...</i>	1	1	1	1	1
Geometric measurement: understand...volume	3	3	3	3	3
<i>Understand volume as an attribute of solid figures...</i>	1	1	1	1	1
<i>Measure volumes by counting unit cubes...</i>	1	1	1	1	1
<i>Relate volume to [multiplication and division]...</i>	1	1	1	1	1
Geometry	4	4	4	4	4
Graph points on the coordinate plane to solve...	2	2	2	2	2
<i>Use [two] perpendicular lines...to define a coordinate...</i>	1	1	1	1	1
<i>Represent... points in the first quadrant...</i>	1	1	1	1	1
Classify two-dimensional figures...on...properties	2	2	2	2	2
<i>[Know category] attributes [apply] to all sub-categories...</i>	1	1	1	1	1
<i>Classify...figures in a hierarchy based on properties</i>	1	1	1	1	1
Total	26	26	26	26	26

Multiple interim assessments built to this design would have different sets of test questions, but with the same emphasis on each of the content categories as on the summative assessment.

Modular interim assessment designs are different, however. Modular designs are intended to focus in on strategically selected subsets of the content standards (typically selected to represent potential

moderate-sized units of instruction). Therefore, modular interim assessment designs are not similar to the summative test design. For example, in a highly simplified approach, each of the five broadest content categories could be selected as the focus for each of five interim assessment modules, giving the following modular interim assessment design of approximately the same length as the mini-summative designs:

Content Category	# of Items on Interim Assessment				
	1	2	3	4	5
Operations & Algebraic Thinking	27				
Write and interpret numerical expressions	18				
<i>Use parentheses, brackets, or braces...</i>	9				
<i>Write simple expressions that record calculations...</i>	9				
Analyze patterns and relationships	9				
<i>Generate...numerical patterns...given rules...</i>	9				
Number & Operations in Base Ten		28			
Understand the place value system		16			
<i>Recognize [digit values increase tenfold when one place... left]</i>		4			
<i>Explain patterns in... when multiplying by powers of 10...</i>		4			
<i>Read, write, and compare decimals to thousandths</i>		4			
<i>Use place value understanding to round decimals to any place</i>		4			
Perform operations...to hundredths		12			
<i>Fluently multiply multi-digit whole numbers...</i>		4			
<i>Find whole-number quotients of whole numbers...</i>		4			
<i>Add, subtract, multiply, and divide decimals to hundredths...</i>		4			
Number & Operations—Fractions			28		
Use equivalent fractions...to add and subtract fractions			8		
<i>Add and subtract fractions with unlike denominators...</i>			4		
<i>Solve [fraction word problems by comparison...]</i>			4		
Apply and extend...multiplication and division			20		
<i>Interpret a fraction [as a division problem]...</i>			4		
<i>[Extend whole number] multiplication to...fractions...</i>			4		
<i>Interpret multiplication as scaling (resizing)...</i>			4		
<i>Solve...problems [with] multiplication of fractions...</i>			4		
<i>[Extend division to involve unit fractions]</i>			4		
Measurement & Data				25	
Convert like measurement units [in the same] system				5	
<i>Convert among different sized measurement units...</i>				5	
Represent and interpret data				5	
<i>Make a line plot to display [data with fractional units]...</i>				5	
Geometric measurement: understand...volume				15	
<i>Understand volume as an attribute of solid figures...</i>				5	
<i>Measure volumes by counting unit cubes...</i>				5	
<i>Relate volume to [multiplication and division]...</i>				5	
Geometry					28
Graph points on the coordinate plane to solve...					14
<i>Use [two] perpendicular lines...to define a coordinate...</i>					7
<i>Represent... points in the first quadrant...</i>					7
Classify two-dimensional figures...on...properties					14
<i>[Know category] attributes [apply] to all sub-categories...</i>					7
<i>Classify...figures in a hierarchy based on properties</i>					7
Total	27	28	28	25	28

The benefit of a modular interim assessment design is that it can provide much more granular and instructionally useful information because there are enough items measuring fine-grained categories

of content to inform broad (not day-to-day) instructional and/or remedial decisions. Another benefit such designs offer is that if districts administer all of the modular interim assessments the time devoted to statewide summative assessment could be considerably reduced.

APPENDIX E: MATRIX SAMPLING TO REDUCE REQUIRED STATE TESTING TIME

This appendix will include additional information about a matrix sampling approach to allow for decreases in required time for state summative assessments if districts administer module-based interim assessments covering all of the content addressed by the state summative assessment.