THE WYOMING COMPREHENSIVE ACCOUNTABILITY FRAMEWORK: PHASE I

Produced for the:

WYOMING SELECT COMMITTEE ON STATEWIDE EDUCATION ACCOUNTABILITY

By

Scott Marion, Ph.D. & Chris Domaleski, Ph.D.

NATIONAL CENTER FOR THE IMPROVEMENT OF EDUCATIONAL ASSESSMENT

January 31, 2012

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Senate File 70, passed during the 2011 Wyoming legislative session, outlined sweeping changes to Wyoming's educational assessment and accountability systems. The legislation specified the creation of a legislative Select Committee on Statewide Educational Accountability and an Advisory Committee to support the Select Committee's work. The Select Committee directed the Legislative Service Office (LSO) to secure the services of the National Center for the Improvement of Educational Assessment (Center for Assessment) to serve as technical consultants to both committees on accountability design and development. The two committees agreed that the first major task—referred to as Phase I in this report—was to create a comprehensive accountability framework so that the major accountability and assessment initiatives work together coherently to best improve Wyoming's educational accountability system. The second aspect of Phase I was to specify the general design of the school accountability system. This report presents the Wyoming comprehensive accountability framework which describes the fundamental elements that must be addressed to design, operationalize, and evaluate a credible and technically defensible school accountability system that supports Wyoming's goals. The framework also addresses the key considerations essential to establishing an educator and student accountability system.

Goals of the System

It is important for the framework to be guided by a well-articulated theory of action. This theory of action specifies the goals, purposes and uses for each accountability system. Additionally, it defines the assumptions, actions, and mechanisms hypothesized to bring about the desired outcomes. Finally, the theory of action should support coherence across multiple accountability initiatives. The first step in developing a theory of action is to specify the goals of the system. The Select and Advisory committees articulated the following as the goals for the Wyoming educational accountability system:

- Have Wyoming become a national educational leader among states
- Have all students leave Wyoming schools "college or career ready"
- Increase the rates at which Wyoming students learn
- Reduce and eventually minimize gaps in achievement in Wyoming
- Improve the quality of teaching and leading in Wyoming
- Maximize fiscal and strategic efficiency of Wyoming education
- Increase credibility of and support for Wyoming public education

The Wyoming School Accountability System

School Accountability Indicators

Drawing from the priorities in the theory of action and aiming to meet the system goals, the following indicators were proposed for the school accountability system.

Category	Recommended Indicators
Achievement	Performance on PAWS reading, mathematics, writing, and science.
Growth	Measure of student progress for reading and mathematics anchored to a standard based on attaining or maintaining proficiency.
Readiness	Status and growth measures on EXPLORE, PLAN, and ACT; graduation rate or index. Additionally, include a broader set of measures for reporting only that includes post-secondary success.
Equity	Additional measure of student progress for non-proficient students only in reading and mathematics. Measure should be anchored to a standard based on attaining proficiency.
Inclusion	Student participation in PAWS, EXPLORE, and ACT.

The Performance Levels: Ratings for Schools

It is one thing to report school performance on each of these indicators, but it is another to summarize the available data into an overall rating for each school. There were extensive discussions with both the Select and Advisory committee about the most meaningful way to report overall performance for each school. Most members of both committees wanted the state to produce an overall rating for each school each year, while others indicated a strong and justifiable preference for avoiding a single rating. The full report provides considerable discussion about the tradeoffs with either approach. The Select Committee recommended producing an overall rating that classifies school performance as follows:

- Exemplary/Exceeding Expectations
- Satisfactory/Meeting Expectations
- Approaching/Partially Meeting Expectations
- Priority Improvement/Not Meeting Expectations

The committees' decision was based on both technical (e.g., a single rating that combines multiple components are more reliable than any individual component) and policy related (e.g., producing outcomes in a manner consistent with policy values will mitigate the risk of misuse). The Select Committee recommended the use of a "decision matrix" as the preferred method for combining the multiple components into a single rating.

In addition to the overall rating, the Select and Advisory committees also recommended producing indicator summaries for at least each of the following categories of school performance:

- Mathematics achievement
- Mathematics growth
- Reading achievement
- Reading growth

- Science achievement
- Writing achievement
- Readiness

Both committees recommended that the state engage in a deliberative "standard setting" process to establish overall levels that are tied to important criteria of performance. This involves generating descriptions of expected overall performance (performance level descriptors) at each of four proposed overall levels outlined above.

Reporting System

The six indicator level subscores will help provide much more "actionable" information than the overall rating, but even that level of information does not contain enough detail to fully inform decisions about supports and program improvement. We argue that it is critical to develop a full reporting system that equips educators, leaders, and stakeholders with ample information at multiple levels. A well-designed and useful reporting system goes beyond static reports and takes advantage of innovations such a dynamic reporting technology and data visualization.

Consequences and Supports

The committees outlined appropriate consequences and supports tied to outcomes in order to promote continuous improvement. The framework presents a multi-tiered system where the overall level triggers a general action, which is further specified according to the performance on the various indicators. In general, schools with higher overall performance are granted greater flexibility and schools with lower performance receive more intensive interventions that correspond to the areas most in need of improvement. A system of supports is critical to accountability system effectiveness and both committees recognize the need to do more design and development in this area.

Developing an Educator Accountability System

The accountability framework also provides an overview of the elements that must be addressed to implement an educator evaluation system. These include: defining purpose and uses, selecting multiple measures, incorporating academic growth, addressing attribution, and quantifying sources of error. This section of the framework provides recommendations for key operational challenges such as defining teacher/leader of record, dealing with missing data, and addressing the challenge of non-tested grades.

Developing a Student Accountability System

In response to the directive in Senate File 70 to review alternatives to the current body of evidence (BOE) system, the framework presents a series of considerations for using end-of-course (EOC) tests to determine if students are eligible for high school graduation. The framework presented recommendation for using a defined process for making critical decisions about the components of such a system. This process should include key stakeholders who address key decisions that include:

- Defining a "Wyoming graduate"
- Clarifying the required knowledge, skills, and dispositions of a Wyoming graduate

- How a set of EOC tests can serve a student accountability system
- Decisions about the many issues related to the development of an EOC assessment system
- The types of support and interventions that must accompany such a student accountability system.

<u>Implications for Standards and Assessments</u>

Standards and assessments are fundamental components of the school, educator, and student accountability system. Therefore, the framework provided an overview of the essential characteristics of a standards and assessment system best poised to support accountability goals. **The Select and Advisory Committees unanimously and strongly recommended that Wyoming formally adopt and implement the Common Core State Standards (CCSS)** because of the standard's strong link to college and career readiness, clear articulation of knowledge and skills across grade levels, and support for comparability across states. In terms of large-scale assessment, the framework discusses the importance of ensuring key criteria such as alignment to the knowledge and skills associated with post-secondary readiness, comparability, reliability, and validity.

Given that SF 70 authorized the use of benchmark computer adaptive testing to measure student longitudinal growth as part of the state accountability system, the framework addresses significant concerns with this approach. Specifically, we recommend not using a test such as Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) or similar benchmark assessments for <u>accountability</u> purposes because:

- it is contrary to the purposes for which the assessments are typically designed,
- of concerns about the technical quality of such assessments, and
- such uses degrade the instructional value of the assessments.

The final section of the framework discusses the critical need for a comprehensive evaluation of the accountability system prior to and following implementation in order to maximize the likelihood that the state's objectives will be met. This evaluation should address the extent to which evidence supports the claims in the theory of action and the degree to which results are reliable and valid. In particular, it is critical to pilot the model in advance of high-stakes accountability uses and study the outcomes to refine system decisions. Moreover, ongoing monitoring and investigation should inform decisions to promote continuous improvement of the system.

SECTION I: BACKGROUND

Introduction

Wyoming Senate File 70 set forth an ambitious agenda to reform the ways in which Wyoming schools, educators, and students are held accountable for academic performance. While this new law will undoubtedly create some implementation challenges, Wyoming has the opportunity to do something few states have done. By enacting such comprehensive accountability legislation, Wyoming has the opportunity to create a coherent educational accountability framework to improve the likelihood of realizing the goal of making Wyoming education the envy of the nation. This coherence will not emerge simply by following the requirements of the legislation. Rather, the State needs a comprehensive accountability framework to describe in much more detail than can and should be presented in legislation the various components of each system—school, educator, and student—and how they fit together to form the overall Wyoming educational accountability system. This document presents this comprehensive accountability framework to guide the development of current and future accountability systems in Wyoming.

The Wyoming legislature enacted this sweeping legislation out of a strong desire to increase the quality and reputation of Wyoming's educational system, to ensure that Wyoming students can compete effectively in the "flat world" of the 21st Century, and to attract and foster economic development in Wyoming. The sweeping accountability legislation was also motivated by a desire to monitor and perhaps improve the financial efficiency of public education in Wyoming. As several members of the legislature questioned, "are we getting the right bang for the considerable number of bucks we are putting into the educational system?" To be clear, legislators were not looking to reduce funding, they simply wanted to make sure that, as responsible public stewards, they were spending the public's money as wisely as possible.

SF 70 was an ambitious piece of legislation that was created under tight timelines as well as other pressures. As such, it is not perfect. In fact, one of the main purposes of this comprehensive framework is to help guide the development of new legislation during the 2012 session based on a luxury of a more deliberative approach followed during the 2011 interim. Therefore, the reader will notice that many recommendations in this report are not perfectly aligned with SF 70 and occasionally are at odds with the language of SF 70.

This comprehensive accountability framework provides an overview of the elements that must be addressed to design, operationalize, and evaluate a credible and technically defensible education accountability system that supports Wyoming's goals. This is particularly important given the broad reach of Senate File 70 and the multiple purposes and uses of assessment and accountability described. The comprehensive framework outlines the fundamental requirements for school and educator accountability with a focus on establishing coherence among all components. The comprehensive accountability framework is organized in five major sections with multiple chapters within each section, as follows:

I. Background
II. Conceptual Foundations
III. The Multiple Accountability Initiatives
IV. Consequences, Support, and Capacity Building

V. Evaluation and other Technical Considerations

This framework was based on recommendations from the Wyoming Select Committee on Statewide Educational Accountability during the 2011 interim. Additionally, this framework benefited from guidance provided by the Advisory Committee to the Select Committee on Statewide Education Accountability that met several times during the interim as well as providing input via email and telephone conference calls. Given the short time frame during the interim and the broad scope of Senate File 70, it is beyond the scope of this document to provide detailed specifications and recommendations for all areas of the framework. The framework presents a broad sketch of the entire system and outlines the steps necessary to further define aspects of the system not explicated here. For example, in the section on student accountability, we discuss some key considerations to help ensure coherence with the full system, but then outline a process by which the specific decisions could be made.

Senate File 70

Senate File 70 created the "Wyoming Accountability in Education Act" and originally set forth a two-phase approach to the development of a comprehensive accountability system. The first phase directed the Wyoming Department of Education to take specific actions relative to an accountability and statewide assessment system. The second phase established the Select Committee on Statewide Education Accountability and an Advisory Committee of education stakeholders to develop a long-term accountability system. In fact, the two phases have essentially been reformulated such that Phase I has focused on the development of a school accountability system, while Phase II looks to the longer term when educator and student accountability systems are included in the larger framework. We summarize the provisions of SF 70 by using the following categories:

- Statewide assessment
- Statewide accountability including required and recommended indicators
- Longitudinal data systems and reporting
- Policies, consequences, and supports

This report does not deal with aspects of SF 70 in this summary that focus on school funding (e.g., *School Finance Recalibration*) or related matters. Additionally, the intent of this section is to provide a brief summary of the provisions of the law. We offer comments about the provisions in the appropriate sections of the report. For example, we discuss assessment issues in Section XI of this report and in doing so, offer comments and recommendations about the assessment provisions in SF 70.

Statewide Assessment

Most urgently, SF 70 required the Wyoming Department of Education (WDE) to eliminate the open-response questions on the PAWS reading and mathematics tests and to use a writing assessment comprised of a single writing prompt to be administered at a time of the year distinct from the NCLB assessments. The legislature also directed WDE to issue a Request for Proposals (RFP) to hire an assessment contractor to implement the requested changes for the school year 2012-2013.

The legislature also directed the State Board of Education (SBE) to develop and implement statewide benchmark adaptive assessments for the 2012-2013 school year to be administered at the district level. Further, the law directed the SBE to use these assessments for evaluating student growth in math and reading in grades K-8. The Advisory Committee recognized the challenges of using the same assessment system for both instructional improvement and accountability as well as the more powerful growth models available for the state summative assessments and, therefore, recommended not using the benchmark adaptive system to fulfill the accountability growth component. This is discussed in more detail in both the school accountability and growth sections later in this report.

SF 70 directed the SBE to "align statewide assessment components" with the accountability system. This recognizes the need to ensure that the assessment system is able to support the requirements and demands of the accountability system. This is discussed in detail in a subsequent section of this report. Additionally, the legislature directed the SBE to consider alternatives to the current body of evidence system including the potential of using statewide end-of-course exams to replace the body of evidence system. Section III discusses this in the context of a student accountability system.

Furthermore, the legislature required the administration of two of ACT's tests. The ACT will be administered to all grade 11 students in reading, English, mathematics, and science, while the EXPLORE will provide information about the performance of eighth graders in the same four content areas, but may be administered in the fall of the ninth grade.

Statewide Accountability

The legislature suggested a two-phase approach to the development of the WY comprehensive accountability system. Phase I directs the WDE to begin reporting the performance of Wyoming schools on a variety of indicators, categorized as achievement (status), college readiness, and growth/improvement, while Phase II authorizes the creation of a Select Committee on Statewide Education Accountability along with an Advisory Committee to support the Select Committee to review the indicators and other aspects of decisions that occurred as part of Phase I. In actuality, Phase I and II have operated concurrently and have been somewhat reconceptualized. While certain assessment aspects of SF 70 have been operating according to schedule, this comprehensive accountability framework is being used to guide the development of all accountability components, but presents a fairly detailed sketch of the school accountability system. Again, we return to this in more detail in subsequent sections of the report. For now, we summarize key aspects of the school accountability provisions as outlined in SF 70.

- Achievement (status)—reading as measured by PAWS in grades 3-8, and 11
- College readiness—percentage of students meeting/exceeding college readiness benchmarks in English, reading, mathematics, and science in the EXPLORE and ACT
- Growth/Improvement—SF 70 specifies a fairly unique approach to measuring improvement of performance of WY schools. The law directed WDE to compute "a combined school score for each core indicator" and measure improvement from year to year, beginning with school year 2011-2012. Since these indicators are computed at the aggregate level, it is more appropriate to call these "improvement" indicators rather than growth, which is often focused at the individual student level. The SF 70 improvement model requires the use of 2010-2011 as the baseline year and then to compare the

> subsequent results such that “positive progress” means that the school achieved a “better score than the year before,” if there was no change from the prior year, the school would be considered “performance level unchanged,” and if the “score declined” from the prior year, it would be called “negative progress.” Through the work of the interim, the Select Committee and the Advisory Committee are recommending a different and more sensitive approach to measuring improvement that is based on evaluating the growth of each individual student. This will be discussed in considerable detail in subsequent sections of this report.

SF 70 directed the Select Committee to design a system of measuring teacher and administrator effectiveness including establishing components of effective teacher and leading. The legislation called for a system to replace the performance evaluation currently in place and to have such a system consider consequences and incentives for improved performance.

Longitudinal data systems and reporting

The legislation directed the WDE to adopt rules and regulations [note: only the SBE can adopt rules] for establishing a system of reporting to include longitudinal data on all aspects of the statewide education accountability system. Importantly, SF 70 directs WDE to create student-teacher links so that assessment results can appropriately and fairly be linked to educators of record.

Policies, consequences, and supports

Senate File 70 directs the SBE to consider consequences, starting in 2013-2014, for failure to meet school accountability targets that focus on the development of improvement plans and then escalate to varying levels of required technical assistance. The law wanted SBE to describe time schedules within which underperforming schools should reasonably be expected to achieve improvement targets. SF 70 also directed the SBE to consider failure to meet target accountability targets in the accreditation process

In terms of educator accountability, SF 70 directed the Select Committee to review merit pay methodologies related to teacher performance measures, including merit-based salary schedules, bonuses, incentive pay and differential staffing practices. This is not a requirement, but is a recommendation for the SBE to consider such consequences/rewards as part of the educator effectiveness system.

The legislation recognized important systematic policy issues that could interact with having SF 70 fulfill its intended goals. First, it authorized the Select Committee to review and make recommendations regarding school district board of trustees training needs. This is an important issue considering that the provisions of SF 70 accountability systems will have significant implications for local boards of education. Finally, SF 70 directed the Select Committee to review the likely effect of current laws on student performance. In other words, if there were existing statutes that might hinder the implementation of one of the accountability systems described herein or otherwise negatively influencing student achievement, the Select Committee should identify and make recommendations to ameliorate potential statutory conflicts.

SECTION II: CONCEPTUAL FOUNDATIONS

Goals and Intended Outcomes

The assessment and accountability system design must be guided by the goals and intended outcomes of the system. These goal statements, which are essentially making explicit the legislative intent, also serve as a foundation for the evaluation of the validity of the policy and associated accountability system. Therefore, a critical activity of both the Select and Advisory Committees was to clearly articulate and come to agree on these goals.

The Select Committee was clear that they wanted to see Wyoming's educational system become recognized as a **national educational leader among states**. The feeling among committee members, supported by data from national assessments, was that while Wyoming's students perform above average on national comparisons, they are still in the middle of the pack. Of course, defining what is meant by a "top educational state" is not easy. States rank order differently on any variety of indicators such as NAEP, ACT, AP, graduation, teacher quality, and countless others. In fact, states often rank order differently on different components of NAEP such as fourth grade reading and eighth grade math, for example. ACT and SAT scores are notoriously tricky to use as indicators of statewide performance, because even if Wyoming were to mandate that all 11th grade students participate in the ACT, it would not be a fair comparison with states that have voluntary participation. There is a notable negative relationship between average state ACT/SAT scores and participation rate, such that the higher participation rates are associated with lower average scores. Therefore, it makes most sense to use fourth and eighth grade state NAEP results as one set of indicators for general educational achievement. Of course, this does not include high school and so, in spite of earlier cautions about using ACT as an indicator, Wyoming's performance could be compared against the other five or six states (e.g., CO, IL, KY, and UT soon) that require census ACT testing of 11th grade students.

Another major goal for Wyoming education expressed by both the Advisory and Select committees was to improve overall levels of student achievement such that **all students leave Wyoming schools "college or career ready**." Of course, there is not universal agreement of what is meant by this term, and both committees recognized the need to further define the separate components of this phrase (college and career). But, both committees clearly expressed the desire to ensure that all Wyoming students leave high school with legitimate options for a career or postsecondary opportunities. The Select committee was particularly insistent that career readiness did not get buried in the rush to define college readiness, because for Wyoming both college and career readiness were equally valued.

If overall achievement rates are going to increase such that all students leave Wyoming schools ready for college or careers, Select and Advisory Committee members recognized that a more immediate goal would be to **increase the rates at which Wyoming students learn** in each academic year. This is essentially a goal that focuses on improving the academic growth that individual students make from year to year in Wyoming schools. Indicators related to this goal can be evaluated using a variety of student longitudinal growth models, which are discussed later in this report.

An important **equity goa**l for Wyoming's educational system is to **reduce and eventually minimize gaps in achievement** among students from historically underperforming student groups. Therefore, a comprehensive accountability system for Wyoming should hold schools accountable for the performance of these groups of students and efforts to reduce such gaps in performance. Additionally, one member of the Select Committee suggested that given the attempts to equalize funding across the state, according to need, we must eliminate the performance gap among school districts. While schools and districts would not be held be accountable for these reductions in gaps among districts, it would be an important goal for the state system as a whole.

While it might go without saying, if all if the goals mentioned above are realized, then the **quality of teaching and leading** in Wyoming schools would have to improve. The two committees recognized the importance of teacher and leader quality as a goal, in and of itself, and declared this to be an important goal of the system in its own right. In thinking through a theory of action (discussed below), improving teaching and leading as a part of both the school and educator accountability systems is a critical stepping stone on the way to improving student learning. It seems obvious that any accountability system should focus on improving the quality of educators in the system, but far too often such systems establish perverse incentives that can actually lead to a decline in educator effectiveness. As part of the coherence principle underlying the development of this comprehensive accountability system, it is critical that the system lead to the positive development of teachers and leaders in Wyoming.

Wyoming lawmakers are proud of the support they have provided to public education, especially over the last 15 years. This is in noticeable contrast to the decimated budgets of public education in many states around the country. On the other hand, as good stewards for the public trust, these same lawmakers are responsible, to the extent they can, for ensuring that public money is well spent. To this end, the Select Committee has stated an efficiency goal for Wyoming education such that the **state is getting an appropriate "bang for its buck**." This should not be read as a desire to scale back the relatively strong funding support experienced by Wyoming schools, rather this goal is simply stating a desire to make sure that all funds allocated for Wyoming to education contribute to the goals outlined above and throughout this document. The legislature, through the Select Committee, has indicated a willingness to spend what it would take to realize these goals, but as responsible lawmakers, would prefer not to spend more than necessary.

Finally, if most or hopefully all of these are realized, the committees would hope to see the **credibility of and support for Wyoming public education increase** among members of the public. This is important for many reasons, but especially if the Wyoming legislature continues its strong support of education, it will be vital that the public recognizes and appreciates the value of this support. Public education is almost always well supported by parents or guardians with students still in school, but as the proportion of the public in this category has shrunk from a high of almost one-third down to less than a quarter of the voting public, it becomes critical that support for education increases its base. As evidence emerges from other states and international locations about the importance of a high quality public education system (actually a P-16 system) for attracting and sustaining business, the policy leaders on both committees recognize that if the educational system improves to the point where it helps improve the business and economic climate, broad-based public support for education will undoubtedly improve.

Guiding Principles

In addition the goals and intended outcomes, accountability system designs benefit by clarity of the key principles used to guide such designs. This comprehensive accountability framework tried to hold true to the following key principles:

- Instructional Core
- Coherence
- Equity
- Transparency
- Support and Improvement
- State-Local Partnership
- Shared Responsibilities

Instructional Core

One of the key design principles in our work has been the "Instructional Core." The instructional core[1] is a set of principles articulated by Richard Elmore and his colleagues that focuses on the relationship among the students, teachers, and meaningful content (and skills). To quote from City, et al (2003):

> *There are only three ways to improve student learning at scale:*
> *You can raise the level of the content that students are taught. You can increase the skill and knowledge that teachers bring to the teaching of that content. And you can increase the level of students' active learning of the content. That's it. Everything else is incidental. That is, everything that's not in the instructional core can only affect student learning and performance by, in some way, influencing what goes on inside the core. Schools don't improve through political and managerial incantation; they improve through the complex and demanding work of teaching and learning (p. 24).*

This is a critical principle and challenges one to think hard about how best to honor this dynamic in the context of designing a large scale accountability system. Nevertheless, the Advisory Committee felt that it was important to maintain a focus on the instructional core throughout the design deliberations.

Coherence

The systems, particularly the school and educator accountability systems must incentivize common and mutually supportive behaviors among teachers and leaders in schools. Wyoming, as a result of SF 70, has a unique opportunity to design school, educator, and student accountability systems all within a short time frame. This will allow Wyoming to develop mutually reinforcing and coherent systems, but this is easier said than done. There are many ways to get tripped up on the way to coherence and the current and subsequent design

[1] City, E. A., Elmore, R. F., Fiarman, S. E, & Teitel, L. (2003). Instructional rounds in education: A network approach to improving teaching and learning. Cambridge, MA: Harvard Educational Press. [see particularly, chapter 1: The Instructional Core].

committees need to continually check design systems within any one of the systems against the likely unintended negative consequences that could occur within that system as well as within the other systems. For example, an indicator for the school accountability system is improved student achievement in reading and mathematics, but if the educator evaluation system was designed such that there was a "zero sum game" where only half or so of the educators in a building could be rated high on the growth indicator, the two systems would be in direct conflict because educators would not have an incentive to work together to improve the performance of the overall school.

Equity

To match the intended outcome of improving the equality of educational opportunities for all Wyoming students, the Advisory Committee recognized the importance of designing the accountability system to support the reduction in gaps of performance/growth for specific groups and individual students. This would play out in terms of a design principle by ensuring that key indicators in the system are disaggregated by specific groups of students, that the accountability metrics are not designed to mask underperforming groups, and the system incentivizes behaviors to promote improved performance of all students in the system.

Transparency

Unfortunately a very simple accountability system is rarely fair and an extremely fair system is rarely simple. Nevertheless, the Advisory Committee urged that the design of the system must be only as complicated as necessary to support the major goals and guiding principles. No matter how complex, the workings of the system should be as transparent as possible such that anyone using the same data set and with appropriate technical understanding could replicate the analyses for any school or the state as a whole. Further, the State must communicate the design and results of the system in ways that can promote an accurate understanding of the system for as many stakeholders as possible.

Support and Improvement

An accountability system can be designed to rate schools or teachers according to some criteria. If that is all that occurred, the accountability would not fulfill the intended goals and outcomes described above. Both the Select and Advisory Committees were clear that the systems should be designed to maximize opportunities to support and improve schools' and educators' performance rather than focus on punitive sanctions. In fact, both groups recognized that it made little differences regarding the accuracy with which the system could label or rank schools, if there was not a parallel system of support, interventions and capacity building also in place. This is discussed in considerable detail later in this document.

State-Local Partnership

Given the strong local control culture in Wyoming and to ensure that districts are encouraged to play their critical role in improving and supporting schools, the systems will be designed to incorporate district expertise and capacity in the accountability design. If the system is to function as intended and realize the goals set forth herein, this cannot be seen solely as a top-down state compliance mandate. Rather, districts and schools will have to be engaged and included as partners in key aspects of the design, implementation, and support associated with the various accountability initiatives if the system will lead to improved outcomes for Wyoming students.

Shared Responsibility

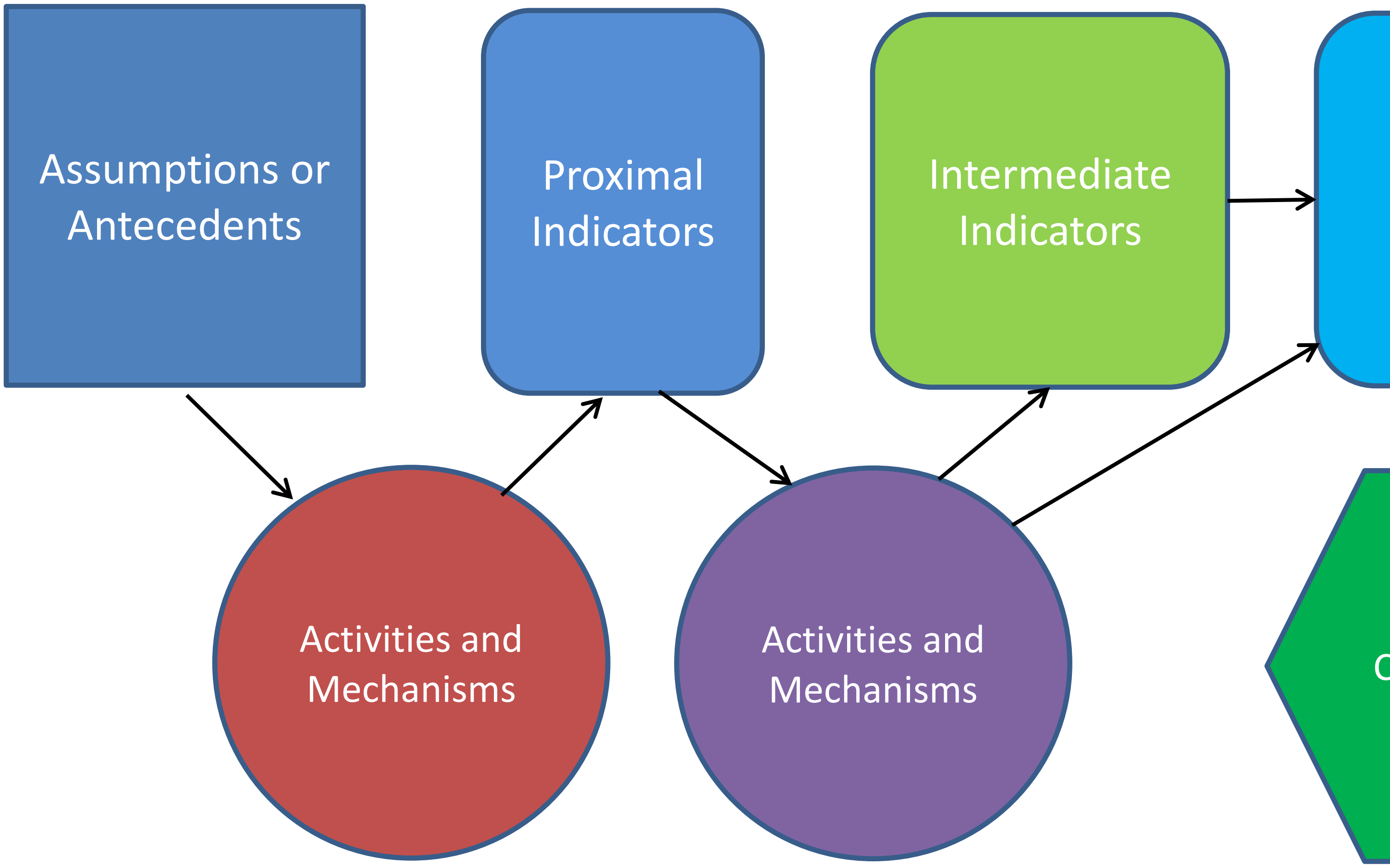
The Advisory Committee recognizes and wants to make clear that the issue of improving Wyoming education is not solely a function of educators or even educational policy makers. Rather, the committee was emphatic that this needs to be a shared responsibility among parents, students, communities, and all policy makers. We use a few examples to illustrate this critical issue. If the school accountability system is going hold high schools accountable for ensuring that its student graduate, the state legislature could support this goal by requiring that students not be eligible to legally drop out of school at least until their 18th birthday. At least one state that has increased the dropout age from 16 to 18 years has seen a noticeable reduction in the dropout rate. A more exaggerated example can be seen in the discrepancy between the penalties for having a truant child compared with getting a ticket for having a dog running at large. A more substantial example involves the investment that would be required if the State was to seriously attempt to address the gap in educational opportunities that are present before students even enter kindergarten. To fully address this issue with universal, high-quality day care and preschool, appropriate nutrition and medical care, along with a host of other opportunities would require a significant policy and fiscal commitment. There is certainly a wealth of evidence to suggest such investments in early childhood health and education is associated with significant long term benefits to both individuals and society. These are just a few examples of how some significant educational challenges can be addressed by both within school initiatives and external policy support.

Theory of Action

A theory of action (TOA) is a useful tool for designing for accountability systems. A TOA explicates the goals of the system, clarifies assumptions supporting or constraining the system, and most importantly explicates the mechanisms by which the various components work together that describe how the system will bring about the desired results. Several researchers (e.g., Bennett, 2010; Marion & Perie, 2009) have employed theories of action as a way to help states and others both design and evaluate complex accountability and assessment systems. A theory of action, drawn from the evaluation literature (e.g., Wholey, 1979), is intended to portray what is essentially a logic or causal model that describes how programs are intended to work. A theory of action lays out the inputs or antecedent conditions, proximal, intermediate and distal outcomes, and importantly describes the mechanisms or processes that specify the logic by which these components are sensibly related.

The general structure for a theory of action is seen below in Figure 1. Following this schematic, we present the foundational principles for the entire system. We then outline the various components of the theory of action for the school accountability systems. Subsequent reports in Phase 2 should provide a theory of action of the educator and student accountability system.

Figure 1. Basic Structure of a Theory of Action.


Assumptions or Antecedents
Proximal Indicators
Intermediate Indicators
Activities and Mechanisms
Activities and Mechanisms
O

<u>Major Goals (Intended Outcomes) of the System</u>

1. Improve overall levels of student achievement such that all students leave Wyoming schools “college or career ready.”
2. Increase the rates at which Wyoming students learn in each academic year (growth).
3. Reduce and eliminate gaps in achievement and especially growth for key subgroups.
4. Improve teacher and leader quality in Wyoming.
5. Increase public credibility and support for Wyoming public education.
6. Increase the “efficiency” of schooling in Wyoming.
7. Have Wyoming viewed as a national education leader among states.

<u>Antecedents</u>

1. Schools are funded at levels adequate to support high levels of student achievement.
2. The learning targets (standards) are clear and support curriculum and instruction.
3. Educators (teachers & leaders) have the knowledge and skills necessary to improve student learning.

4. The state summative assessments in ELA and mathematics provide technically defensible student scores for reporting a “status” (proficiency) measure related to the state content standards.
5. The state summative assessments in ELA and mathematics provide technically defensible student scores for calculating the growth in student performance across consecutive school years. The school accountability system supports a collective vision of school improvement and responsibility.
6. Key stakeholders agree that the school accountability system represents a broad set of indicators necessary for characterizing school quality, while focusing on those indicators most likely to leverage positive change.
7. Schools and districts have the capacity to support the data collection and improvement efforts related to school accountability.
8. WDE has the capacity to implement and support the school accountability system including working with schools to improve their performance over time.

Proximal indicators (numbers) and mechanisms (bullets)

1. Measuring and reporting student longitudinal growth provides information that educators use to judge the quality (effectiveness) of educational programs.
    - ➢ Educators and other stakeholders will use this information to fine-tune, alter, and/or eliminate specific programs/interventions to focus on those with the greatest likelihood of producing gains in student learning.
    - ➢ Having access to high quality information on student progress will allow educators to more easily develop cultures of data use for making educational decisions.
2. Measuring and reporting student longitudinal growth provides information for students, parents, and other key stakeholders to more accurately judge the progress each student is making for each school year.
    - ➢ Parents and others will advocate for more effective educational programs and interventions for their students.
    - ➢ Students will receive information that will enable them to better monitor their own progress.
3. District-selected interim assessments fully aligned to WY standards and/or CCSS and administered at least multiple times throughout the school year are used to monitor student learning throughout the school year.
    - ➢ Teachers and others use the interim assessment results to monitor and adjust the instructional programs for students.

Intermediate indicators (numbers) and mechanisms (bullets)

1. Clear and actionable assessment/accountability reports accurately portray schools in terms of achievement (status), student longitudinal growth, and other key indicators (e.g., graduation rates, college/career readiness).
    - ➢ Data are used to improve the quality of interventions and programs at Wyoming schools.
    - ➢ The assessment system, accountability calculations, and reporting systems provide information for school leaders to support and improve the quality of teaching.

2. The data and decisions from the school accountability system contribute to local educator evaluation systems in ways that allow excellence to be recognized and collaboration is encouraged.

Distal indicators

1. The average teacher and leader quality statewide improves and the variance at the lower ends of quality is reduced.
2. There is an increase in high quality applicants for open teaching positions.
3. Students grow at rates that lead to increased levels of college and career readiness compared to current rates.
4. Student achievement will improve statewide as evidenced by increases on state assessments, NAEP, and related assessments.

Consequences (intended and unintended)

1. The system is designed in such a way as to maximize the likelihood of the distal indicators being fulfilled.
2. Schools that do not meet prescribed state accountability standards are subject to increasing levels of actions including filing school improvement plans, working with a "distinguished educator," replacing the school leader, and/or other consequences as determined by the State School Board.
3. Schools that excel on school accountability indicators may be afforded certain flexibility such as freedom from certain WDE or other requirements.
4. The accountability system does not lead to a narrowing of the curriculum or other meaningful opportunities for students.
5. The accountability system does not lead to Wyoming teachers leaving the state for other teaching opportunities

We note that there is a strong coherence between "Theory of Improvement" embedded in the Wyoming Funding and the theory of action presented here as well as in the accountability framework designed to create incentives for continuous improvement of student performance in Wyoming. The "Theory of Improvement" in the Wyoming Funding system stresses the foundational point that core instruction is the prime route to improved student performance. In addition the Theory adds several elements that together should operate to improve instructional practice. These include: very small class sizes; high teacher salaries; school-based instructional coaches; and all the resources needed for an ongoing professional development system. In addition, the Theory of Improvement embedded in the Wyoming Funding system includes multiple resources for both Tier 2 and Tier 3 (in a Response to Intervention model) extra help for students struggling to meet performance standards. These include: adequate numbers of professional staff for one-to-one as well as small group tutoring and other interventions; extended day programs; summer programs; and one hundred percent reimbursement for all special education costs.

SECTION III: THE MULTIPLE ACCOUNTABILITY INITIATIVES

This section of the report presents information and recommendations for developing school, educator, and student accountability system. As noted earlier, we provide considerably more details on the development of and recommendations for the school accountability system since that has been the focus of the Phase I efforts. We then outline key considerations and recommendations for processes to develop educator and student accountability systems. As discussed above, a key principle guiding the development of this section of the report was an intention to create a coherent approach to educational accountability such that the important goals set forth earlier might best be achieved.

As part of development a comprehensive accountability and support system, the Advisory Committee worked from a theory of action focused on continuous improvement of the system. As part of these discussions, the committee recommended clarifying the differences among data collection, reporting, and accountability and supported an approach whereby more data were collected and reported than might be used as accountability indicators. The intent is not to create a "data dump," but to collect information on targeted areas that could be useful to schools for improving the performance on the accountability indicators. For example, graduation rate will be a key indicator for the school accountability system, but the advisory committee recommended collecting data and reporting results on indicators such a 9th grade credit accumulation because of its strong relationship to dropping out of school. The reader may question why we are not including 9th credit accumulation in the accountability system if it is such a good indicator, but the committee recognized quickly the highly corruptible nature of such an indicator if used for high stakes accountability.

School Accountability Framework

Introduction

In this section we describe the overall framework for the school accountability system, possible indicators that will likely comprise the core components addressed in the school accountability system, and some initial thoughts about how the various indicators may be combined to create overall determinations. This will be followed in subsequent sections by a more in-depth treatment of design issues.

Indicators

The building blocks of an accountability system are the indicators or measures that produce information about school performance. Indicators serve at least two critical functions in an accountability system. First, the selected measures signal and, hopefully, promote the valued behaviors for school leaders and educators. For example, if it is desirable to increase achievement in mathematics, including performance on mathematics assessments should encourage schools to focus on mathematics instruction. In this manner, the identified indicators ***serve as a policy lever to promote desired actions***. It should be clear, then, that the identification of indicators must be closely linked with the theory of action for the accountability system. Second, ***indicators contribute to overall measures or classifications of school performance.*** Accordingly, measures should be selected that capture an important component of school quality linked to the intended use. For example, if the desire is to identify schools that are ‘failing’ and should be considered for restructuring, indicators must be selected that provide information well-suited to differentiate and classify schools that meet minimum performance expectations from those that do not.

Naturally, to the extent that indictors are used to influence high-stakes accountability outcomes, they must be reliable and trustworthy. There will almost certainly be dimensions of school quality that are important to capture but are too variable or corruptible to be used for high-stakes purposes. For example, policy makers may agree that ‘parent engagement’ is an important dimension of school quality, but in the absence of a suitably meaningful and standardized method for measuring this component, it would be unwise to include the indicator for high-stakes decisions. This is not to suggest that schools should not attempt to measure or even, in some circumstances, publicly report outcomes. Rather, our caveat pertains to use of ‘soft’ measures in high-stakes decision making.

In selecting and defining indicators there are a number of additional considerations that should be carefully weighed. We can regard these considerations as being related to 1) the number of indicators 2) type of information produced and 3) unit of analysis. With respect to the number of indicators, it is certainly desirable to include varied information to better understand and account for the many factors that define school effectiveness. Generally speaking, the inclusion of multiple measures bolsters the validity of the outcomes. On the other hand, too many elements may make the model complicated to understand and burdensome to implement. Taken to the extreme, such an approach could be regarded as simply a ‘data dump’ where it is difficult to detect the signal through the noise. There is a real risk that by including too much, policy makers will lose sight of what is most important. For this reason, we recommend that the system be built around indicators that reflect the most prominent values in Wyoming’s theory of action.

The second consideration is related to the measure or type of information one elicits from the indicators. For example, when considering assessment results one might use a scale score or classification with respect to an identified standard (e.g. basic, proficient, advanced) which can be aggregated and reported as ‘percent proficient.’ The latter approach carries the advantage of being straightforward and easy to interpret. However this masks degrees of difference within performance levels, which is conveyed with a scale score. Similarly, when working with outcome measures, such as graduation, one can produce a broad measure, such as graduation rate, which simply reports the percentage of students in a cohort who achieved this outcome in a set period of time. Alternately, a more granular approach to including outcome indicators can be adopted that provides detailed information, but may add to the complexity of the system.

Finally, it is important to consider the unit of analysis for the selected indicators. Critically, decisions about unit of analysis should match the goals and priorities of the system. Because an important outcome is to ensure equity of opportunity and achievement, it is essential to track indicators for groups of students for whom equity concerns are most important (e.g. students with disabilities, English language learners, economically disadvantaged students etc.). For example, consider test performance as an indicator. This can be reported as percent proficient and aggregated to the subgroup, grade, school, district, or state level (or some combination thereof). If the decision is to report for relatively small units, such as subgroups within schools, there must be a high degree of confidence that the student information system supports this and an understanding that some units may be very small and data may be highly variable and ill-suited to support inferences. Finally, the sheer volume of information produced will make the design of clear, coherent reports more challenging. On the other hand, if the system is based on a higher level of analysis, this will likely be more straightforward to operationalize and report and better suited to support inferences. However, this higher level of aggregation may mask important information for policy makers.

In selecting and defining indicators, the overall goal is to create a balanced model that is suitably 'granular' to provide specific actionable information but sufficiently robust to support meaningful claims about school performance. Additionally, the model should be simple and transparent enough to be easily understood and implemented.

Based on the requirements of SF 70 and the feedback received from the Select and Advisory committees, we propose the following indicator categories.

A. **Achievement** – How do students perform on state assessments designed to measure proficiency on Wyoming state standards?
B. **Growth** – Are students demonstrating acceptable progress with respect to performance on state standards?
C. **Readiness** – Do students graduate college and career ready?
D. **Equity** – Are the lowest performing students attaining proficiency or demonstrating acceptable progress toward proficiency?
E. **Inclusion –** Are all students participating in the accountability system?

In the sections that follow, suggestions for identifying specific indicators to be included as well as advice for including these components in the accountability system are presented. For added clarity, design illustrations are presented to aid in conceptualizing alternatives. However, these illustrations should not be regarded as exhaustive or proscriptive, rather they are intended to help bring shape to ideas in order to better evaluate options to promote intended policy objectives.

Achievement

Achievement refers to indicators that provide information about student academic performance with respect to Wyoming state standards. At a minimum, Senate File 70 proscribes that the system address "core indicators of student performance" to include reading as measured by PAWS – grades 3-8, and 11. In addition to reading, we recommend inclusion of PAWS mathematics results in the accountability system.

The inclusion of science and writing was a matter of some debate in the Advisory Committee meetings. While committee members endorsed the importance of promoting achievement in science and writing there was some concern that the current assessments were not well suited to promote the desired outcomes and should have little to no influence in the accountability model. However, the Select Committee was clear that if the goal is to promote science and writing instruction, these two subjects must be included in the model.

Furthermore, an “alternate assessment” for the students with the most significant cognitive disabilities must be included (according to the Individuals with Disabilities Act of 1997) in each grade/ content area for which a general assessment is incorporated in the achievement calculation.   This ensures that schools are accountable for the performance of all students.

Achievement Design Illustrations

As noted earlier, there are number of options for how to include achievement information in accountability systems. A common alternative is to use percent of students meeting a target performance standard – typically proficiency or level 3. While this measure is fairly course, it is conceptually clear to stake holders and prioritizes a valued outcome.

Given that there are multiple grades and content areas, one way to accomplish this is to simply compute the ratio of all proficient students across all grades and content areas at the school, and divide this by all examinees as depicted in Table 1.

Table 1. Illustration of Combined Proficiency Calculation.

Number of Math Examinees	200	Number Math Proficient	160	Percent Math Proficient	80%	Total Proficient (330/405)	81.5%
Number of Reading Examinees	205	Number Reading Proficient	170	Percent Reading Proficient	83%		

The resulting percentage can then be adjusted by a factor to determine the overall weight or influenced in the model. For example, if proficiency is intended to be expressed on a scale from 0 to 300, multiplying the result from Table 1 (.815) by 300 will produce a metric that ranges from 0 (no students proficient) to 300 (all students proficient). In the example depicted in Table 1 a school would receive 245 of 300 points.

There are a number of possible variations on this approach. One variation is to weight one content area test more or less than another. For example, if science were included and one desired that science results account for only 20% of the outcome, math and reading could each be adjusted to contribute 35% each to the overall outcome, with 10% of the weight coming from writing and the remaining influence (20%) would come from science.

Another variation is to create a performance index such that schools get some ‘credit’ for students in level 2 – rather than an ‘all-or-nothing’ measure. This can be accomplished by creating a ratio such that student scoring at levels 2, 3, 4 on the state assessment and those scoring at levels 3 and 4 only are divided by all examinees.   This figure would be multiplied by 150 (half of the maximum value of the scale) to get a total score out of 300. By so doing, schools essentially receive half of a credit for students who score at the basic level and a full credit for students who score at the proficient or advanced level (see Table 2).

Table **2Table 2: Illustration of Index with 'Partial Credit' for Basic Performance**

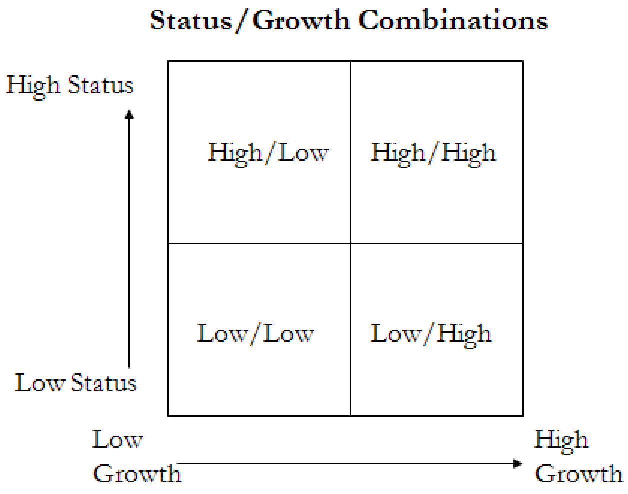
Performance Level	N	Number Basic or Above	Number Proficient or Above	Calculation	Result
Below Basic	40	160	105	$\left(\frac{160+105}{200}\right)150$	199 out of 300
Basic	55				
Proficient	85				
Advanced	20				
Total	200				

Growth

The achievement category is based on 'status' indicators, which show how students are performing relative to a criterion (proficiency) at a single point in time. However, it is also important to include growth, which measures change in performance for the same student or cohort of students over time. Examining the combination of growth and status performance for schools provides a much richer picture of school quality than either component in isolation.

Figure 2 shows 4 possible outcomes for schools taking into account both status and growth. Naturally, the most prized result is for schools to be in the top right quadrant, where most or all students are proficient on state tests and all students are growing at a high rate. The converse of this is shown in the bottom left quadrant in which relatively low percentages of students are proficient and the growth rate is also low – an obviously undesirable outcome. Including growth also helps identify and give credit to schools in which proficiency may be low but students are growing at an exceptionally high rate (bottom right quadrant). On the other hand, it's important to understand which schools have traditionally high performing students, but show relatively low or no growth (top left quadrant). This may describe a school with affluent, historically high achieving students who are languishing.

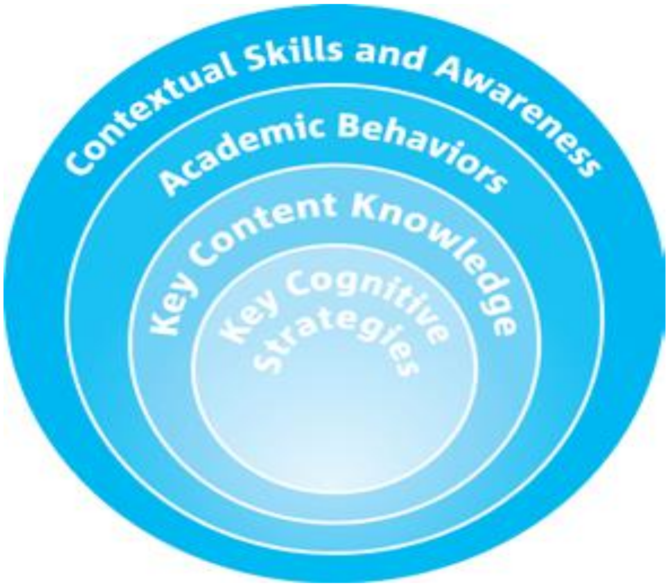
Figure 2: Status and Growth Combinations



There are many promising approaches to measuring and including growth in education accountability systems. Due to the scope and complexity of this issue, we address this topic separately in the next section of this document.

Readiness

In an accountability system that prioritizes college and career readiness it is important to include indicators that signal that a student is prepared to be successful in college or a career or is 'on-track' to meet this expectation. There are numerous potential indicators for this category, particularly when one considers that 'readiness' is a multi-faceted dimension that goes beyond academic performance and includes such characteristics as academic behaviors. David Conley (2005) and his colleagues at the University of Oregon have provided a powerful framework for thinking about college readiness. This framework is depicted in the following graphic and described below.



- ✓ Key Cognitive Strategies are also known as "habits of mind" and include skills such as inquisitiveness, persistence, and intellectual openness.
- ✓ Key Content Knowledge is broken into overarching types of knowledge such writing and the ability to conduct research and core academic knowledge that includes much of the focus of high school learning, such a mathematics, language arts, science, and social studies.
- ✓ Academic Behaviors are critically important skills for independent learners to possess and include such things as self awareness, meta-cognitive, and self-regulation.
- ✓ Contextual Skills are often referred to as "college knowledge" and include knowing how to navigate oneself around college system and deal with such things as financial aid, applications, enrollment, and other details that can easily sideline otherwise "ready" students.

This invites consideration of ‘non-traditional’ measures which can provide a much broader view of readiness, but also presents challenges related to lack of standardization or corruptibility of measures. For this reason, we suggest distinguishing between current standardized readiness measures that are suitable for contributing to school accountability classifications versus those measures for which valid and reliable data are not yet available that should be reported but not used for high-stakes decision making at this time. The types of measures that fit this latter category include many of the following, among other potential indicators, and the Advisory committee recommends that Wyoming explore the possibility of collecting and reporting the results for several of these indicators as data become available and are deemed trustworthy. Further, the committee endorses including such indicators in the reporting system before they are used in an accountability context.

- Course completion/ success
    - Enrollment and/or performance in AP/IB or other ‘advanced’ courses
    - Participation in joint-enrollment or other post secondary courses at the secondary level
- Qualitative data (e.g. survey data of attitudes, academic habits etc.)
- Attainment of career/ industry certifications
- Achievement of post-secondary outcomes
    - Enrollment in credit bearing courses
    - Attainment of qualifying career, enrollment in the military etc.

While these are not common to school accountability models and may be difficult to track, it can be argued that they provide valuable information to evaluate the fundamental claim that students are on track to or have exited high school ready for college and/or the workforce. It should be noted that these are *preliminary* ideas discussed by the Advisory Committee and we suggest additional exploration with higher education and workforce leaders to better understand what is feasible (e.g. data capabilities) and appropriate to include.

Alternatively, we suggest two categories of indicators that we believe are promising for inclusion in accountability determinations: academic performance and graduation rate.

Academic performance refers to achievement on tests that are explicitly linked to college or career readiness. The ACT suite, as it is typically called, includes the following three assessments. Both the EXPLORE and ACT are specifically cited in SF70.

- EXPLORE: measure of progress toward college readiness, typically administered in grade 9 (but reflects performance through grade 8).
- PLAN: measure of progress toward college readiness, typically administered in grade 10 (but reflects performance through grade 9).
- ACT: measure of attainment of knowledge and skills associated with college readiness, typically administered in grade 11.

Graduation rate provides an indication of student outcomes at the completion of high school. Naturally, the most desirable outcome is for students to graduate on-time with a diploma that certifies the student is ready to succeed in college or the workforce. Other less desirable outcomes are also possible such as a GED or certificate of attendance.

Readiness Design Illustration

A straightforward approach for including EXPLORE, PLAN, and ACT scores in the accountability system could correspond to the method previously described for achievement (i.e. PAWS) indicators. However, in lieu of proficiency, the primary criterion becomes the percent of students meeting an identified readiness benchmark. One simply multiplies the percent of students meeting the benchmark by the selected maximum value of the category. Another approach would be to create an index for ACT scores that would be based on key benchmark. For example, schools could be awarded 50 points for each student scoring at the entry level benchmark into credit-bearing classes for Wyoming community colleges (e.g., 18), 100 points for scoring at the national average, 125 for scoring at the important college ready benchmark of 24, and 150 points for students scoring at or above a score of 27 or 28. This is only an example. The actual benchmark scores and point values should be recommended by the Advisory Committee after gathering appropriate information from higher education and other stakeholders. This index could be computed on the ACT composite score, but might be more useful if computed at the individual test level (math, reading, science, language).

There was some concern that incorporating ACT into the accountability is simply adding another “status” measure that is correlated with student and school socioeconomic status. The Advisory Committee was interested in exploring the use of either or both improvement and growth measures to provide a way for less advantaged schools to do well on this metric. For example, schools could be evaluated on how much their ACT index or indices change every year of over a three year period. Similarly, schools could be evaluated on how much student performance improves as the students move from the EXPLORE to the PLAN and then to the ACT. This could be the fairest measure of high schools’ contribution to readiness, since it takes into account where students start in this domain.

While graduation rate can be similarly incorporated into the accountability system, it may desirable to consider multiple levels of performance. To accomplish this, an index can be created that awards points in proportion to the value of the outcome in year 4 as illustrated in

Table **3**. The score for this component is simply is the average of all student outcomes for the high school.

Table 3: Example of Graduation Index

Student Result	Points
Diploma with completion of required college/ career ready course work	100
Other diploma	85
GED	50
Continued enrollment (no outcome)	25
Certificate of attendance	25
Dropout	0

The data in the table are illustrative only, the actual categories and point values would be determined based on Wyoming's goals and policy priorities. Importantly, both the categories and values should be defined by bringing together a broad-based group of Wyoming education leaders and stakeholders to define priorities.

One additional factor to consider is that students may graduate in more than four years. While this is less favorable, there may be important reasons to account for and incentivize this result in an accountability model. One approach to account for this is to award incentive or bonus points for outcomes in subsequent years. For example, a student who maintains enrollment in year four but does not earn an outcome receives the corresponding points in the index (25). If the following year the student earns a GED they get a portion of these points (e.g. 10%) added to the index value for their school. The incentive points are then averaged for all students with delayed outcomes and added as a 'bonus' to the index.

Equity

Another category that should be addressed in a comprehensive accountability system is the extent to which *all* groups of students are achieving success. In the best case, not only will schools improve achievement overall, but they will also erase what are often persistent and sizeable gaps in performance between highest and lowest performing student groups.

There are at least two key questions to consider in evaluating alternatives for equity measures.

1. Which group(s) should be the primary focus for equity?
2. What equity outcomes are most important to promote?

Equity groups can be defined based on one more demographic factors (e.g. ethnic group, economically disadvantaged status, students with disabilities). Or, it is possible to combine multiple groups in a single subgroup. By so doing, schools that otherwise would have too few students in any one group to produce a determination will be included in equity outcomes. Additionally, the larger group size will produce more stable results.

Another way to define focal groups for equity, which we believe is the most promising alternative, is to determine membership based on performance as opposed to demographic factors. For example, the group is defined as students who fail to meet proficiency on state tests. This approach ensures that schools focus on improving outcomes for all students who are low performing.

A second consideration is determining the equity outcomes that should be promoted in the accountability system. In keeping with the values inherent in SF70 and expressed by the Advisory committee, we propose that the expectation for students below proficient is to demonstrate satisfactory academic progress or growth to proficiency. Specifically, we recommend producing a separate growth measure for non-proficient students that is meaningfully linked to attaining or maintain proficiency. This will exert substantial influence on the results for schools and explicitly communicate progress of low performing students, rather than masking outcomes in summary data. Moreover, this will reward schools making the most progress with low performing students and penalize schools making the least progress. In the subsequent section on growth, we will provide more details regarding this proposed approach.

Inclusion

Finally, schools must be accountable for including all students in accountability determinations. This helps insure that results are not manipulated by excluding low performing students. This can be addressed in a straightforward manner by reporting participation rates for all indicators and setting a very high minimum threshold, such as 95%. However, it is reasonable to include results in performance determinations for only those students who were present at the school for the full academic year. These aspects are typically handled in the 'business rules' for operationalizing the system, which is otherwise beyond the scope of this document.

Growth

In this section we provide an in-depth discussion of using growth in a comprehensive accountability system, with a detailed illustration of design alternatives using Student Growth Percentiles (SGP).

Growth Alternatives

During the Advisory and Select Committee meetings, members were introduced to and discussed a variety of approaches to measuring academic growth. Although classification schemes have limitations (most notably: they are not mutually exclusive), four general categories of growth were presented to aid in conceptual clarity: categorical, gain score, value-added, and normative. These approaches and the prominent advantages/ limitations of each are summarized in Table 4.

Table 4: Overview of Growth Alternatives

Method	Description	Answers what question?	Advantages	Limitations
Categorical	A measure of the change in performance level category from time 1 to time 2	Did the student advance or decline performance levels?	-Straightforward to understand and implement - Clear relationship to status	-Insensitive to large growth or overly sensitive to small growth -Influenced by test properties -Not well suited for very high and very low performing students
Gain Score	The difference between scores between time 1 and time 2	What is the magnitude of student growth?	-Straightforward to understand and implement - Results on a familiar scale with known relationship to status	-Requires vertical scale - There are technical concerns with vertical scales - Magnitude of growth cannot be interpreted the same for all students
Value-Added	Regression based approach that controls for multiple variables to determine the difference between actual and predicted growth	To what degree was the student's performance higher or lower than that of similar students?	- Accounts for multiple factors that influence growth -Provides a definition of 'typical growth' based on similar students -Expectations are adjusted based on abilities and characteristics	-More complex to implement -Including background variables can be controversial -No 'built-in' relationship to status, but growth targets can account for this
Normative	Regression based measure that conditions current achievement on prior achievement to describe performance relative to students with identical prior achievement	To what degree is performance higher or lower than expectations, based on students with similar academic history?	-Provides a familiar basis to interpret performance – the percentile -Provides a definition of 'typical growth' -Expectations are adjusted for students of various abilities	-More complex to implement -No 'built-in' relationship to status, but growth targets can account for this

As should be evident, there is no single correct approach to growth or method that stands-out as the 'gold-standard.' The decision regarding which analytic approach should be adopted should first be considered in context to the purpose for measuring growth and the desired model characteristics. In the best case, the selected model should produce outcomes that are reliable and valid for the intended uses and produce results that are clear and easily understood by stakeholders. Additionally, the model should be practically feasible to implement and maintain.

Given that alternative analytic approaches and model specifications will produce different growth results, it stands to reason that a policy-based decision regarding which model is most suitable for Wyoming should also be based on the extent to which a given model most reliably detects schools/ classes judged to be high or low performing. In other words, all else being equal (e.g. equally technically viable and equally operationally manageable) the model that produces results most in sync with Wyoming's definition of quality should be prioritized. For example, if the state heavily values academic growth for the lowest achieving students (e.g. those below proficient) then a model that is more sensitive to detecting progress for students below standard should be prioritized.

Growth Expectations

Another critical decision related to implementing growth measures for accountability purposes is establishing growth standards. More plainly, 'how much growth is good-enough?'

Broadly, approaches to identifying growth standards can be characterized as either norm-referenced or criterion-referenced. A norm-referenced approach compares student achievement to a statistically derived expectation, such as the mean performance for students with similar prior achievement. Growth that exceeds this predicted value is judged to be 'good,' whereas a growth rate below statistical expectation is regarded as 'bad.'

Alternatively, criterion-referenced growth standards establish a specific target outcome. For example, requiring students who are not proficient to grow at a rate such that they achieve proficiency in a set amount of time is a criterion referenced approach.

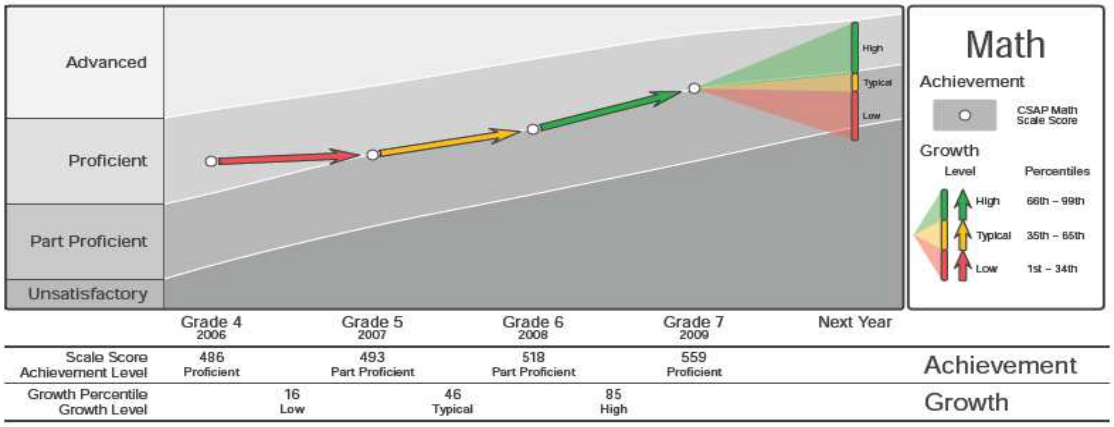
Each approach has advantages and limitations. Setting a norm-referenced expectation is useful for identifying comparably high or low growth. Indeed, it seems intuitively reasonable to describe valued growth as that which is significantly higher than that of similar students. However, a limitation is that some students who grow at very high rates relative to their peers may not achieve proficiency in a reasonable amount of time. A criterion-referenced standard resolves this potential 'growth to nowhere' problem, but raises a new issue: some students may be so far below standard that even at exceptionally high rates of growth the student will not achieve proficiency in a reasonable time frame. Particularly when growth is used for accountability purposes, this can create a condition where some classes or schools are uniformly disadvantaged. Conversely, very high performing classes or schools could exhibit little or no growth and meet standard.

An appreciation of this tension between criterion and norm-referenced growth leads to the conclusion that neither approach alone is adequate. Therefore, we recommend blending the two in the accountability system. In the subsequent section, we introduce the Student Growth Percentile (SGP) as a normative measure of growth and then describe how it can be evaluated with respect to a meaningful criterion.

Student Growth Percentiles

The Student Growth Percentile (Betebenner, 2009) is a regression based measure of growth that works by conditioning current achievement on prior achievement and describing performance relative to other students with identical prior achievement histories. This provides a familiar basis to interpret performance – the percentile, which indicates the probability of that outcome given the student's starting point. This can be used to gauge whether or not the student's growth was atypically high or low as depicted in Figure 3.

Figure 3: Sample Student SGP Report



	Grade 4 2006		Grade 5 2007		Grade 6 2008		Grade 7 2009	Next Year	
Scale Score Achievement Level	486 Proficient		493 Part Proficient		518 Part Proficient		559 Proficient		Achievement
Growth Percentile Growth Level		16 Low		46 Typical		85 High			Growth

In Figure 3, an SGP was calculated for each year this student was enrolled (from grade 4 to grade 5, from grade 5 to grade 6 and from grade 6 to grade 7). At the right of Figure 3, low, typical and high growth is classified by broad percentile ranges. For this hypothetical student, the growth percentile of 16 is classified as "low" and as illustrated in Figure 3, the student's performance dips from being classified as Level 3 in grade 4 to becoming a Level 2 in 2007. In subsequent years, this student's SGP increases to the point that he or she is re-classified as a proficient student in grade 7.

These individual SGPs can be aggregated to evaluate growth taking place at the classroom, school, or district level. Since the median is a more appropriate measure to use with percentiles than the mean, the median growth percentile is typically reported by states using SGPs to quantify average growth taking place at aggregated levels.

Catch-Up/ Keep-Up Growth

As noted previously, establishing appropriate growth expectations for accountability should incorporate both norm and criterion referenced standards. The Catch-Up/ Keep-Up (CUKU) method, initially developed for the Colorado growth model, provides a rich example for how this can be accomplished[2].

[2] See http://www.cde.state.co.us/Accountability/Downloads/GrowthStandardsAccountability.pdf for more information about the application of norm and criterion referenced growth in Colorado.

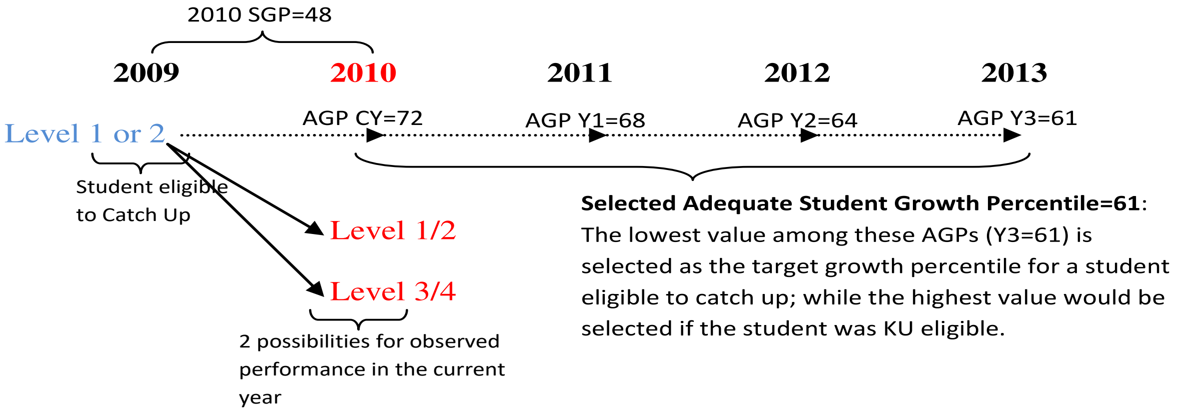
With the CUKU metric two distinct groups of students are evaluated together: students who scored below proficient (Level 1 and 2 students) and proficient students (Level 3 and 4 students) in the prior year. A student is placed in the 'Catch-Up' category if his or her prior year score is below proficient. 'Keep-Up' students are those that were proficient or higher in the prior year.

Then, for the current year and three future years an adequate growth percentile (AGP) is calculated. Each AGP sets the projected growth percentile required for a student to cross the cut score threshold from below proficient to Level 3 in a given grade for the projected year. Each student has an individual AGP that applies specifically to him or her.

From the four AGPs, a single value is selected as an overall representation of a student's needed growth. For a student in the CU category, the selected target is the *lowest* AGP value from among the current or projected year AGPs. This represents the growth a student needs to cross the threshold into the Proficient category or Level 3 at any point in the current year or the next three years. For students in the KU category, the selected target is taken from the *highest* AGP target value. This means a successful Keep Up student cannot fall below Level 3 in the current year or next three years.

Figure 4 shows how the selected AGP is derived for a CU student scoring a Level 1 or 2 in 2009. During the current 2010 school year, the student can either be in Level 1 or 2 or in Level 3 or 4. In this hypothetical case, the amount of growth needed to move from Level 2 to Level 3 decreases from 2010 to 2013. The minimum value selected to represent the AGP for this student is the SGP of 61 from year 3. In essence, the AGP value for a given CU student quantifies how much that student should have progressed in the current year in order to attain proficiency in the future.

Figure 4: Illustration of the Catch-Up/ Keep-Up Method

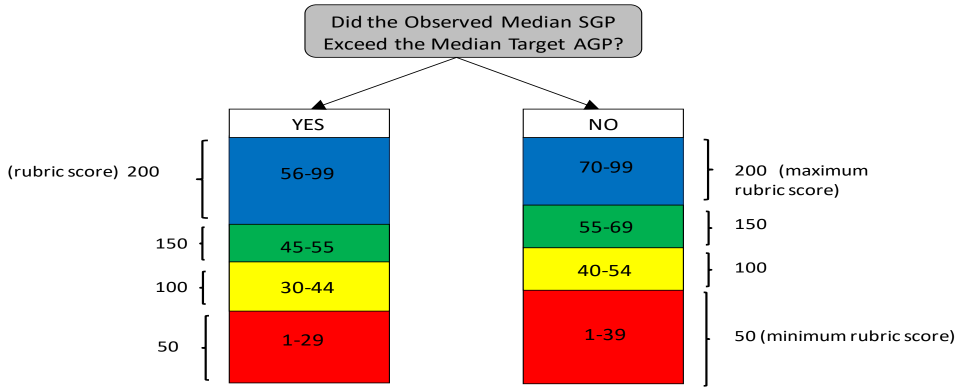

2010 SGP=48
2009
2010
2011
2012
2013
Level 1 or 2
AGP CY=72
AGP Y1=68
AGP Y2=64
AGP Y3=61
Student eligible to Catch Up
Level 1/2
Level 3/4
2 possibilities for observed performance in the current year
Selected Adequate Student Growth Percentile=61: The lowest value among these AGPs (Y3=61) is selected as the target growth percentile for a student eligible to catch up; while the highest value would be selected if the student was KU eligible.

Growth Design Illustration

There are a number of promising alternatives for incorporating SGPs into Wyoming's accountability system. The approaches illustrated in this section evaluate the SGPs relative to proficiency targets based on the Adequate Growth Percentile (AGP) defined in the preceding section. As explained, an AGP is calculated for every student. For a student who scored below proficient in the prior year, the AGP target represents the growth percentile needed for that student to become proficient in one of four years considered.   For a student who scored proficient in the prior year, the AGP represents the growth percentile needed to maintain proficiency across the four years considered.

In the same way that the median is taken across the individual SGP values to evaluate "average" growth taking place at a school, the median can be taken across the unique AGP target calculated for every student depending on whether that student is a below proficient or already proficient in the prior year. Figure 5 illustrates how growth can be evaluated at the school level by using these two pieces of information (median SGP and median AGP) and then evaluating whether the median SGP achieved falls under one of four rubric point categories.

Figure 5: Illustration of Rubric Scores for Schools Meeting or Not Meeting AGP Target


Did the Observed Median SGP Exceed the Median Target AGP?
YES
NO
(rubric score) 200
56-99
150
45-55
100
30-44
50
1-29
70-99
200 (maximum rubric score)
55-69
150
40-54
100
1-39
50 (minimum rubric score)

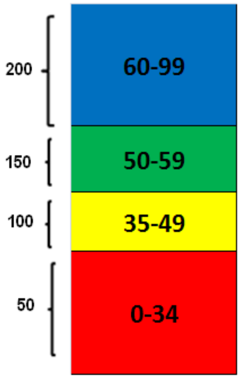
In Figure 5, for a given school, the median SGP is first compared to the median AGP. If the observed median SGP for the school in a given year meets or exceeds the median target (AGP), then the scoring rubric to the left is used to assign rubric points to the median SGP achieved by the school. If the school's median SGP is below the median target AGP, then the scoring rubric to the right is used to assign rubric points to the median SGP achieved by the school. For example, if a school had an observed median SGP of 65 and a median target AGP of 45, this school would be awarded maximum points of 200 on growth as indicated by the scoring rubric to the left. The rubric cut-scores set for schools that meet or exceed their median AGPs are lower than the cut-scores for schools that do not meet their median AGPs since these schools are populated with students who are either largely on track to meeting proficiency or growing at a sufficient rate to maintain their proficient status. The rubric cut-scores for schools that do not meet their median AGPs are set at a higher bar, since these schools need to grow at higher rates in order to move all their students towards proficiency.

Alternatively, a more simplified method for producing school growth scores could be implemented by removing the AGP component from the school evaluation of growth and using a single rubric to assign a school growth score[3]. Simply, the schools median SGP is evaluated against one rubric to determine the growth score. If this approach is desired, it is important to identify rubric values and growth ranges that meaningfully correspond with attainment of desired

[3] However, we would recommend continuing to report AGP at the student level.

achievement outcomes. For example, analyses should be conducted to determine what percent of non-proficient students who score in the highest growth category achieve proficiency in 3 or fewer years[4]. Figure 6 depicts an example of this single rubric approach.

Figure 6: Illustration of Single Rubric to Determine School Growth Outcomes



Growth and Equity

In the previous section, we introduced the idea that the growth component of the school accountability system could be used to support Wyoming's equity values. In other words, growth measures could be used to determine if the lowest performing students were demonstrating adequate progress.

One way this can be accomplished is to compute growth outcomes twice: once for the whole school and again for students below proficient – the target equity group. As described previously, this provides a substantial incentive for schools to focus on improving the performance of low achieving students and substantially rewards those schools that are successful.

There are a number of ways to accomplish this. One approach is to used the same rubric(s) but apply a different scale to reflect the desired weight (e.g. 200 points for whole school and 100 points for non-proficient students). Additionally, non-proficient students could be counted once in each group (i.e. double counted), which places a strong emphasis on equity, or growth could be calculated separately for proficient versus non-proficient students. Finally, a decision regarding which content areas should be included and how much each should be weighted must be considered.

Design Decisions

One of the most critical decisions in the design of any accountability system is determining how the various indicators will be summarized and reported. Essentially all research and evaluation (accountability systems are one type of evaluation) endeavors involve some form of "data

[4] The Center has conducted analyses in another state revealing that meaningful growth targets can be established with a single rubric, which closely correspond with results from the CUKU approach.

reduction" whereby results are summarized to some degree or another. In other words, we rarely collect and report raw data to stakeholders, rather it is summarized and reported in some manner. The challenge is determining which data to summarize and when to stop summarizing. For example, few stakeholders will question computing either a mean scale score or some other summary statistic (e.g., percent proficient) for the reading assessment in a particular classroom. But even this level of aggregation masks other important considerations such as the degree of variability in the students' scores. Going further, we suggest that few stakeholders would question summarizing the achievement results for a given content area for a given year at a school; however some might have concerns about the meaningfulness of combining such results across content areas to produce an overall achievement measure. As with most of the other design decisions, there is not a single correct approach. Rather, the aggregation decisions need to reflect the values and the intended uses of the results. This section of the report outlines some of the aggregation and reporting issues for Wyoming's school accountability system and proposes a framework that links closely the overall performance levels with specific consequences and supports.

First, we must make clear that this discussion is aligned with previous decisions about having a very detailed reporting structure associated with the various Wyoming accountability systems. In terms of the school accountability system, the intent is to report on each of the indicators with enough specificity to inform decisions and subsequent improvement actions. Further, we recommend designing a reporting structure whereby school personnel have access to much finer grained reports than those produced for parents and other members of the public.

Single or Multiple Ratings

There has been considerable discussion among Select and Advisory Committee members about the ultimate level of aggregation for the school accountability indicators. While not necessarily evenly split, there are two main positions. The first involves producing an omnibus rating for each school, while the second position would have multiple ratings for each school, although there is not yet agreement about the specifics of such multiple ratings. While intended purposes and target audiences must inform the aggregation decisions, we discuss the potential advantages and disadvantages associated with these major reporting decisions.

A major advantage of the single overall rating is its simplicity. If the meanings of the ratings are well understood, it could be a very efficient way to communicate, at least at a surface level, information about the overall quality of schools in Wyoming. Of course, the challenge is finding global performance descriptors that accurately convey meaning about the multiple indicators. The ability to have some control over the "message" is another important advantage of using a single overall rating. This advantage depends largely on one's belief about whether someone (e.g., the media, realtors, or other stakeholders) will find a way to create their own aggregate rating, whether or not it is done by the state. If one believes that there is a reasonable probability of somebody or some group creating and publishing their own overall school rating, then one might want an overall rating as part of the system so that the rating accurately reflects the design choices of Wyoming's policy leaders and advisors. On the other hand, if one either believes there is a low probability of such an occurrence or that it can be dealt with once it occurs, they will not necessarily view the single overall rating as an advantage, at least for the ability to help control the message. Finally, while the validity of an overall rating might be questioned (discussed below), previous research and experience indicates that the overall rating will—all things being equal—be more reliable (e.g., consistent across years) than ratings of individual indicators.

The disadvantages of combining the various indicators into a single overall rating are numerous and are essentially the inverse of the advantages. The risk of combing all indicators into a single rating, while apparently simple, may in fact be too blunt to convey sufficiently nuanced information about school quality. Further, while a single rating might do a fair job at distinguishing the highest and lowest performing schools, it might not be very effective at providing a fair and accurate picture of schools in the middle. An example discussed at a Select Committee meeting is that if growth and achievement were weighted about equally, two schools could get very similar ratings even if one had very high growth and low achievement while the other school had the opposite pattern. There was concern that such a rating would mask very important differences among schools. A similar concern arose when considering variability in performance across the content areas. The potential advantage of "controlling the message" has disadvantages as well. Many recognize that the State will not be able to control all potential users and uses and trying to do so by producing a single rating for each school could be seen as falling into the "two wrongs do not make a right" trap.

If there are disadvantages to producing a single rating, that implies there are advantages to producing multiple scores/ratings for each school. The first major challenge of reporting multiple scores or ratings is the need to decide and agree on the type of reporting that should occur. Of course, this needs to be driven by purposes and uses, but there might always be a tension between how much or how little to aggregate. For example, some have suggested reporting separately by content areas (e.g., math, reading, and science), but combining growth and achievement within content area. Others have suggested combining across content areas, but reporting two overall scores, one for growth and another for achievement. To play out this example further, one can easily make the case for reporting growth and achievement separately for each of the content areas, and even by grade levels as well. The point here is that once we move away from simply reporting individual student scores, we have agreed to aggregate. The question then is how far do we continue to aggregate to find the right balance between summary and information?

The educational psychology literature is quite clear that task-specific feedback is much more effective at leading to improvements in performance than general feedback. Therefore, the more fine-grained the reporting, the more likely it is that the accountability system reports and other information will lead to improvements in student learning. To be fair, simply reporting two or three scores (growth and achievement or content area scores) is probably not specific enough to qualify as "task-specific" feedback. This brings us back to the need to have a very detailed reporting system so school personnel will have information available on which to act. However, public reports do not need to present such fine-grained information. On the other hand, reporting summary information in multiple categories, such as growth and achievement by content area could provide a much more nuanced view of school quality than a single omnibus rating. Further, this could be a useful public information activity by educating the public that quality in education is not as simple as "thumbs up" or "thumbs down." Another potential benefit of reporting information either by content area or by content area and growth/achievement is that it can help school leaders address complacent teachers who might be able to "hide" behind an omnibus compensatory rating. For example, a school with highly effective mathematics teachers may get an adequate overall rating even though the language arts teachers are only performing at a mediocre level. A more discrete reporting system would help shine a light on both the strong and weak areas. Of course, one could take care of such cases in an omnibus rating system by not using a simple compensatory system, but requiring some minimum level of acceptable performance in all relevant areas to receive an acceptable overall rating.

Recommendation: Trying to thread the needle

As can be seen, there are tradeoffs with either approach. The advantages of the single rating point out the disadvantages of multiple ratings and vice versa. We are concerned that reporting only two (growth and achievement) or three (reading, math, science) scores/ratings for each school does not go far enough to address the concerns associated with aggregating all indicators to a single rating. Therefore, we see the choice as being between producing a single overall score/rating and producing ratings in ***at least*** the following seven categories[5]:

- ✓ Mathematics achievement
- ✓ Mathematics growth
- ✓ Reading achievement
- ✓ Reading growth
- ✓ Science achievement
- ✓ Writing achievement
- ✓ Readiness

We make this suggestion because knowing that achievement and especially growth can vary considerably across content areas, we do not think that simply reporting two ratings (e.g., growth and achievement) offers noticeable advantages over a single rating. We go further to recommend that a single rating can be produced along with the more discrete ratings suggested above and that such a system can help meet multiple needs of the system. The single rating is undoubtedly what will get published in newspapers and other summary outlets, but if reports are carefully designed, we would hope that the finer grained information would get reported as well.

Performance level descriptors (PLD) and Standard Setting

One of the reasons for reporting a single, overall rating certainly relates to the reliability issue discussed above. A single, largely compensatory rating will be more reliable that any one of the five or six ratings closer to the indicator level. This greater reliability has important implications for establishing cutscores separating the various levels of performance, especially if the goal is to have at least three or four levels. If there is insufficient reliability, it can often play out as problems with classification consistency. That is, low reliability around the cutscores will lead to schools changing categories for no reason other than the uncertainty associated with the system. Therefore, it will be important to have a reasonably high degree of confidence in the overall classification for a school. If there is a reliable overall rating for each school, then it is less critical that each of the finer-grained reporting categories to have similarly high levels of reliability. This is not advocating low reliability, but simply suggesting the higher reliability of the composite can "protect" the lower reliability of the finer categories.

[5] Note: Not all indicators would be available for each school and grade span. For example, readiness would not be reported for elementary schools.

The Select Committee indicated an interest in establishing four levels of overall performance, but there was no discussion about the number of levels that should be set on the finer categorizations. There is a compelling argument to establish the same number of levels on the component parts as the overall levels, but there is also a compelling argument for using a different number of levels. If the same levels are used for all reported categories, it might make communication easier, but it can also lead to confusion. There is always the risk with using the same levels for each major indicator and the overall level that stakeholders will think they can simply average across the major indicators to arrive at the overall score. While this could be true, it likely will not be the case because of differential weighting and other factors. Therefore, we recommend that four performance categories are used for the overall rating, while three are used for each of the major indicator reporting categories. We elaborate on this below, focusing first on the overall level.

We recommend that the State engage in a deliberative standard setting process to establish overall levels that are tied to important criteria of performance. This involves generating descriptions of expected overall performance (performance level descriptors) at each of the four levels and then evaluating accountability system data (in the initial implementation/pilot year) to essentially match overall school scores to these descriptors. This process will result in recommended scores that mark the boundaries between any two levels (cutscores). These recommended cutscores should then be brought to the State Board of Education for approval. We offer recommended levels and initial descriptions for the four overall performance levels:

- **Exemplary/Exceeding Expectations**: Schools in this category, which is reserved for schools considered models of performance, have demonstrated high growth in all applicable content areas, have average to high levels of achievement (proficiency rates), and have high performance on graduation rates and other readiness indicators (if applicable).
- **Satisfactory/Meeting Expectations**: Schools in this category have demonstrated either high levels of growth or high levels of achievement in all content areas and are meeting state targets for readiness indicators.
- **Approaching/Partially Meeting Expectations**: Schools in this category have demonstrated either acceptable levels of growth or acceptable levels of achievement in some, but not all content areas. Schools in the "approaching" category may demonstrate average or lower performance on graduation or other readiness indicators.
- **Priority Improvement/Not Meeting Expectations**: This category is reserved for schools with unacceptable performance on many or most indicators. Schools in the priority improvement category typically have low levels of achievement in all content areas and demonstrate low to average growth in the relevant content areas and fall short of expectations on graduation and other readiness indicators (if applicable).

We recognize that these category names and descriptors will evolve, but argue that that if the state wants to incentivize improvements in the overall state educational system, the highest performance category should be reserved for schools that are truly demonstrating high levels of performance. Similarly, the priority improvement schools, perhaps a slightly larger group than those in the exemplary category, should be reserved for those schools where the State will direct intensive capacity-building resources, which is described in more detail below. All of these performance categories will be intricately linked to expected actions on the part of the school, district, and state. These actions may be termed "consequences," but given the continuous improvement orientation of the Advisory and Select committees, consequences are all designed from an improvement orientation. In spite of the potential usefulness of this overall categorization, the Advisory committee contends that it is too blunt of an instrument to direct improvement actions appropriately. Therefore, before discussing potential consequences, we turn to the establishment of performance levels on the indicator categories.

The following seven major indicators previously are used as a starting point for thinking about reporting at a finer grained level than the single overall level:

- ✓ Mathematics achievement
- ✓ Mathematics growth
- ✓ Reading achievement
- ✓ Reading growth
- ✓ Science achievement
- ✓ Writing achievement
- ✓ Readiness

The major categories could easily be expanded as the number and type of indicators in the school accountability system expand, but these categories represent a good starting point. As noted above, we suggest that each of these major indicators be categorized into three performance levels to both avoid some potential interpretation problems, but to also recognize that the reliability associated with individual indicators might not be high enough to justifiably support the establishment of four distinguishable performance categories. Therefore, we recommend using, at least as a starting point, three levels of performance for these indicators and that the cutscores should be established normatively such as: exceeding the state average, average performance, and below the state average. This is especially useful for the growth measures, but we argue that it can be useful for the status measures as well. However, we would not be opposed to incorporating some criterion-referencing into the establishment of these levels as well. For example, one may want to require that for a school to be considered "average" on the readiness indicator, they should have a minimum requirement of at least a 75% graduation rate. Again, this is just an example to demonstrate how cutscores on these indicators could be established largely normatively but can also include some important thresholds.

Approaches for Combing Multiple Indicators

There are at least four approaches to combining multiple indicators to yield a single outcome: *compensatory*, *conjunctive*, *disjunctive,* and *profile* methods. Compensatory means that higher performance in one measure may offset or compensate for lower performance on another measure. Conjunctive means that acceptable performance must be achieved for every measure. Disjunctive means that performance must be acceptable on at least one measure. A profile refers to a defined pattern of performance that is judged to be satisfactory, unsatisfactory, or equivalent. A profile approach is often operationalized using a matrix to combine indicators for making judgments.

A compensatory approach recognizes that some degree of variability in performance across indicators may be expected. Such an approach has a higher degree of reliability because the overall decision is based on multiple indicators evaluated more holistically. Moreover, reliability improves because random error in multiple measures tends to cancel. Conjunctive decisions are less reliable because errors accumulate across multiple judgments meaning a school might fail to meet standards due to the least reliable measure. However, this approach may be desirable when it is important to assure that a school does not fall below established standards on any one criterion. A disjunctive method is desirable when any one component is viewed as adequate assurance the school has met expectations. Finally, profiles are useful especially when there are certain patterns that can be described that reflect valued performance that are not easily captured, usually because the combinations of criteria are judged to be not equivalent.

These approaches should not be regarded as mutually exclusive. It is possible, for example, to combine aspects of compensatory and conjunctive ‘rules’ to arrive at a final result. An example of this is a rule that requires both 95% participation AND a minimum score on an index that combines status and growth in order to pass. Requiring schools to meet both participation and a minimum performance level is conjunctive; however, an index that combines both status and growth is compensatory.

Matrix Design Illustration

Using six of the indicators we have introduced in this section, we will illustrate an example for how the indicators can be reported at various levels and then combined to support an overall classification. By so doing, we do not propose that this should represent the entirety of information produced by the system. Rather, we seek to illustrate a design alternative with a manageable number of indicators that should figure prominently in the accountability system.

As illustrated previously in this document, growth, achievement, and readiness (for the present example: graduation rate) can be can be expressed a number of ways. For example, we can report achievement as simply percent proficient or on a scale with a desired range. However, regardless of the metric we can ‘collapse’ the outcome into three categories. Here, we will use the following: *Below the Standard, Meeting the Standard, and Exceeding the Standard.*

Taking into account each content area, this produces six performance categories as depicted in Table 5 below, which would be explicitly reported for each school.

Table 5: Illustration of Reported Performance Categories

Math	Reading	Science[6]	Readiness
Achievement Level	Achievement Level	Achievement Level	Performance Level
Growth Level	Growth Level		

It is possible further collapse this information into an overall score by content area, such as a math performance level that accounts for the combination of achievement and readiness. Alternately, the information can be combined by indicator category, such an achievement score that accounts for the influence of math, reading, and science. There are multiple ways to accomplish this, but perhaps the most straightforward would be to produce an overall proficiency

[6] Growth is not calculated for science because it is tested only once each in elementary, middle, and high school.

rating for achievement and a mean score for growth and apply standards to these values to produce a single performance level for each indicator. It is certainly possible to weight one content area more than another to prioritize a policy value. In any case, the result would be a single performance level for each indicator class: achievement, growth, and readiness.

Using these three level ratings for each of three indicators, a decision table can be produced, as shown in Table 6 that indicates how the combinations of ratings work to provide an overall school classification as: *Exemplary/ Exceeding Expectations, Satisfactory/ Meeting Expectations, Approaching/ Partially Meeting Expectations, and Priority Improvement/ Not Meeting Expectations*.

The shaded cells shows the various level on each indicator class and the bold text in the non-shaded cells shows the overall school classification. The actual classification levels are simply illustrative and many other combinations are possible to reflect the values of Wyoming policy makers.

Table 6: Illustration of Decision Table for Performance Indicators

		Achievement Below	Achievement Meeting	Achievement Exceeding
Readiness Level Below Standards	Growth Below	**Priority**	**Priority**	**Approaching**
	Growth Meeting	**Priority**	**Approaching**	**Approaching**
	Growth Exceeding	**Approaching**	**Approaching**	**Satisfactory**
		Achievement Below	Achievement Meeting	Achievement Exceeding
Readiness Level Meeting Standards	Growth Below	**Priority**	**Approaching**	**Satisfactory**
	Growth Meeting	**Approaching**	**Satisfactory**	**Satisfactory**
	Growth Exceeding	**Satisfactory**	**Satisfactory**	**Exemplary**
		Achievement Below	Achievement Meeting	Achievement Exceeding
Readiness Level Exceeding Standards	Growth Below	**Approaching**	**Satisfactory**	**Satisfactory**
	Growth Meeting	**Satisfactory**	**Satisfactory**	**Exemplary**
	Growth Exceeding	**Satisfactory**	**Exemplary**	**Exemplary**

A strong advantage of using a decision matrix to evaluate performance is the ability to apply specific policy-based criteria to all cells, especially the ‘off-diagonal’ cells. When cells ‘agree’ (e.g. growth, achievement, and readiness are all below standard) the decision of a final classification is usually uncontroversial. However, there may be a policy rationale for evaluating one combination of levels as different from another if they are based on dissimilar indicators. In this manner, policy makers may desire to privilege growth, achievement, or readiness.

Compensatory Design Illustration

Using the indicators we have introduced in this framework, we will also illustrate an example of combining achievement, growth, and readiness using a compensatory approach.

As shown previously in this document, achievement can be expressed as a scale based on proficiency rate. In the most straightforward approach, the percent proficient across all grades and content areas is multiplied by 300 to obtain a scale that ranges from 0 to 300 (e.g. .75 x 300 = 225).

Growth, as shown previously, can also be expressed on a scale with a maximum value of 300. This comes from two components: whole school and non-proficient students. In each case, the median SGP is evaluated against a rubric that awards up to 200 points for the whole school and up to 100 points for growth of the non-proficient students for a total of 300[7].

We also introduced two components for readiness: a graduation index and performance on readiness assessments (i.e. ACT and EXPLORE). We can conceive of a readiness scale with a maximum value of 150 at the high school level, where 100 points is derived from the graduation index and 50 points from assessment performance, calculated as the percent meeting ACT benchmark performance multiplied by 50 (e.g. .80 x 50 = 40). The sum of these values produces an overall readiness index for high schools. To keep the maximum value of points available for all schools the same, high school achievement points could be reduced to 150. By so doing, 'status' measures (i.e. test performance and graduation outcome) would carry the same weight in the calculation in contrast to growth.

Including readiness scores for middle schools represents a unique challenge that should be examined separately. The EXPLORE test is given statewide and provides an 'on-track' college readiness measure for 8th grade students. Conceivably, performance on EXPLORE could be incorporated in the model similar to the ACT to produce a readiness value of up to 50 points for middle schools and achievement could be reduced a corresponding amount (from 300 to 250) to keep the overall values consistent. However, because the EXPLORE test is given in grade 9, a process would need to be developed to associate these values with the schools in which the students were enrolled as 8th graders. This would also create a data 'lag' (which, to be fair, may exist for other indicators).

In Figure 7 and Figure 8, we illustrate two examples of a hypothetical point structure for elementary and high schools, incorporating the elements and values described.

[7] Several variations on this approach are possible, including distinguishing between proficient and non-proficient students (to avoid double counting) and changing the weights (e.g. 150 for each component)

Figure 7: Illustration of Hypothetical Elementary School Point Structure

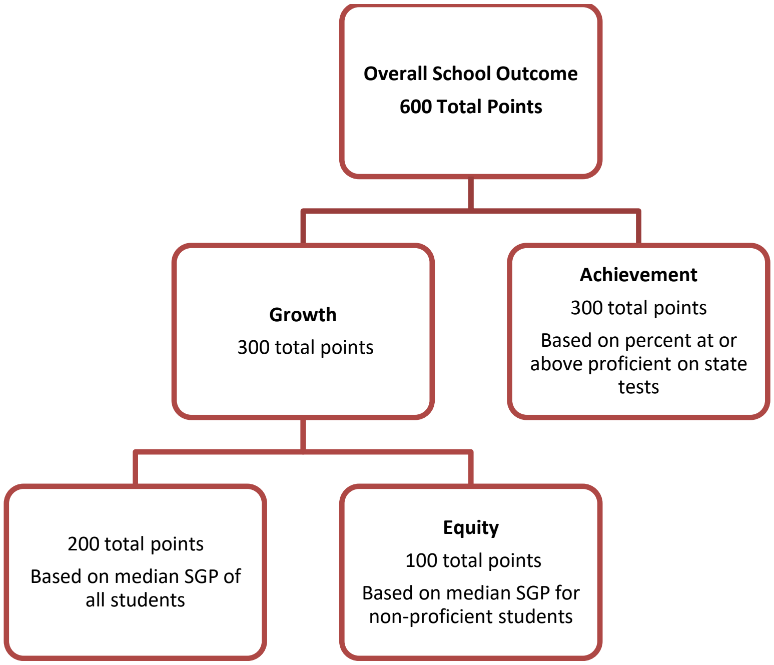
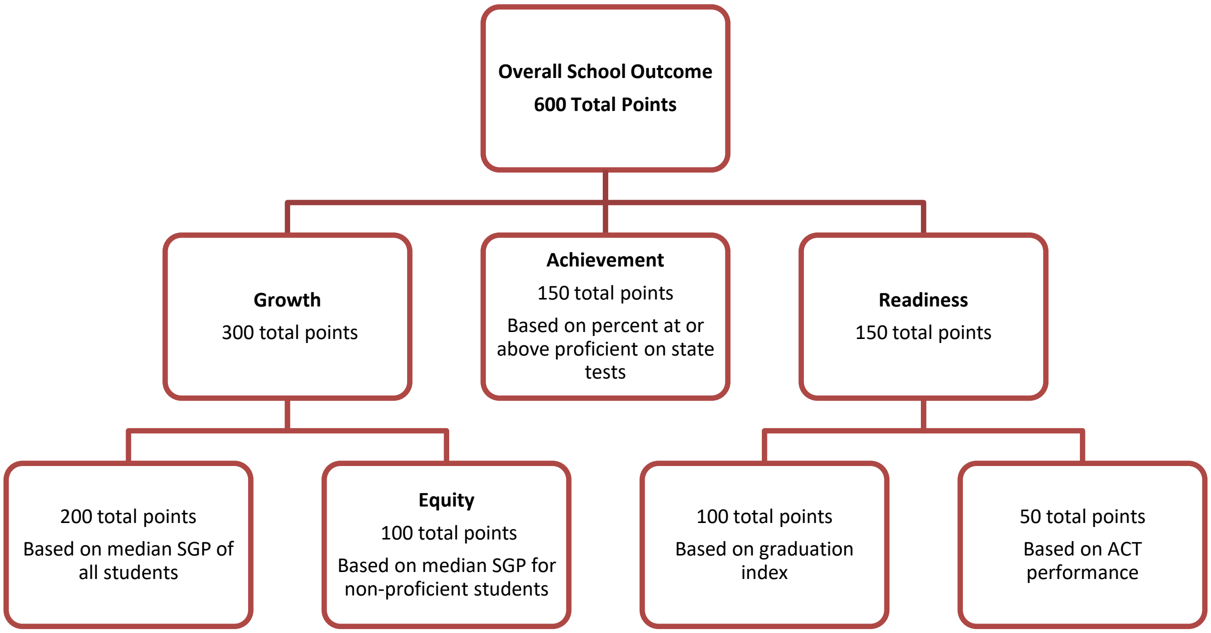

Overall School Outcome
600 Total Points
Growth
300 total points
Achievement
300 total points
Based on percent at or above proficient on state tests
200 total points
Based on median SGP of all students
Equity
100 total points
Based on median SGP for non-proficient students

Figure 8: Illustration of Hypothetical High School Point Structure



The design example portrays a model in which each element exerts influence on the outcome in proportion to the number of points assigned to that component – in this case, achievement and growth are equally valued. Evaluation of school performance is in reference to a target score or threshold on the overall score (e.g. 500 out of 600 to achieve the highest classification.) Schools that score lower on achievement can offset this performance by demonstrating higher growth. Conversely, less growth is required of schools that are already strong in achievement. This illustrates the compensatory nature of the model. *Importantly, the weight of each component and the selection of thresholds are key policy decisions that influence school outcomes.*

It cannot be overemphasized that the values used in this section and previous sections of this document are intended to be purely illustrative to make the ideas presented more clear by example. The actual rubric and scale values should be carefully considered to reflect policy values and modeled to examine impact.

Professional Judgment and Appeals

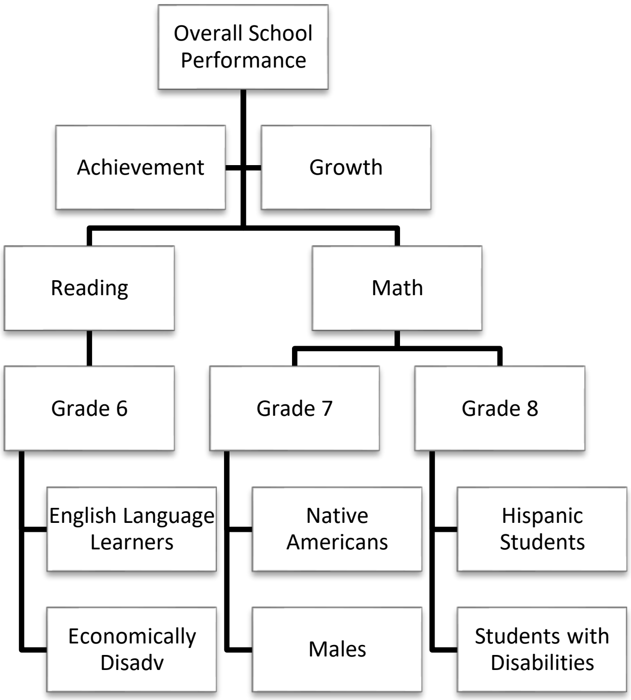
The Select Committee insisted on incorporating professional judgment into any approach for aggregating indicators and producing yearly determinations for schools. While either of the two approaches for combining indicators discussed above—the decision matrix or the compensatory design illustration—must incorporate professional judgment in determining values and weights, the decision matrix approach lends itself to a much more obvious incorporation of policy and professional judgments than seemingly more quantitative approaches. The Select Committee recommended including a professional judgment review of the results each year, particularly for those schools with an “off-diagonal” pattern where the results of the different indicators paint different pictures of school performance. Further, both committees recommended using an appeal process to allow schools to appeal their ratings, first to their local board and then perhaps to a state-convened panel of peers to allow for another layer of professional judgment.

Reporting

As discussed previously, combining content areas or indicators into a single classification has the advantage of being clear to stakeholders and can guard against potentially irresponsible attempts to produce a summary outcome. However, these high level outcomes run the risk of masking important characteristics of school quality. To be sure, more detailed information is needed to inform decisions about supports and program improvement. For this reason it is important to develop a reporting system that equips educators, leaders, and stakeholders with ample information to support a variety of uses.

We envision information available by indicator, by content area, by grade, and by subgroup. However, as depicted Figure 9 below, which is a very small slice of the full range of information that could be produced, it is easy for extant reports to overwhelm stakeholders and serve only as a ‘data dump.’

Figure 9: Sample of Selected Reporting Levels

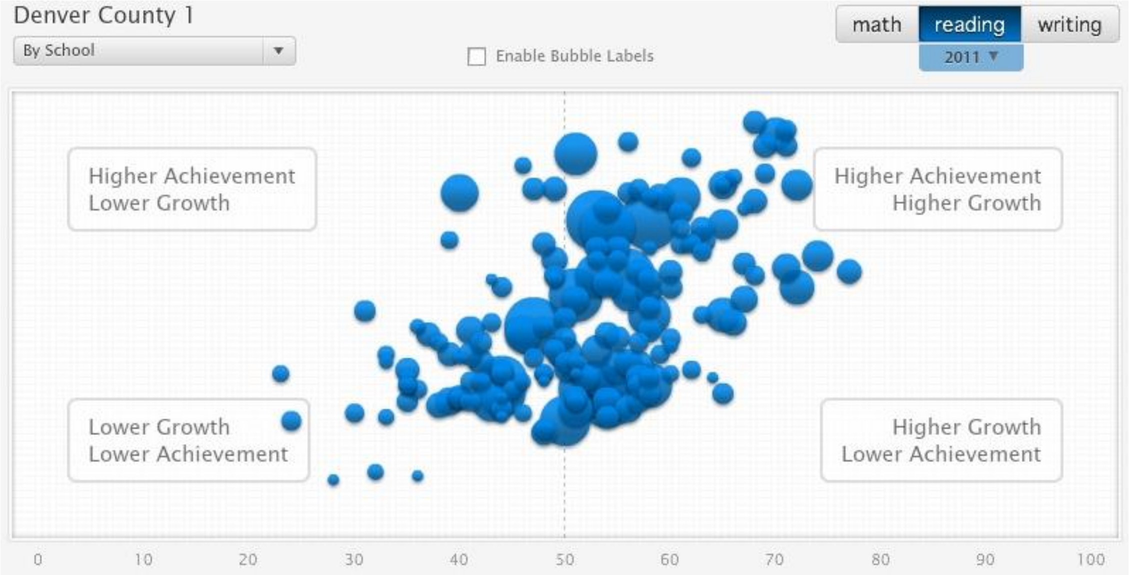


A well-designed and useful reporting system goes beyond static reports and takes advantage of technological innovations. In general, reports should be accessible to stakeholders to ensure that those closest to the classroom have the information needed to inform instructional decisions. Moreover, the reports should be accompanied by adequate interpretative information. Such information should describe the meaning of and precision of the outcomes and clearly indicate uses and interpretations that are supported. Supplemental information may enhance the utility of reports, such as comparative information from similar schools or longitudinal trends.

States that are on the forefront in innovative reporting practices are taking advantage of both dynamic reporting technology (e.g. interactive data tables) and data visualization (e.g. graphs and plots). One such state is Colorado, who employs a system termed SchoolView[8]. In this system, not only can stakeholders access a variety of ‘conventional’ information, such as summaries of state assessment results, but users can produce and manipulate customized reports. It starts with the ability to customize the interface by role (e.g. parent, educator, or administrator). Then, users can access a wealth of information, such as plots of growth (median SGP) by status (percent proficient) and school size. Figure 10 provides an example of this display.

[8] See http://www.schoolview.org/index.asp for more information including access to dynamic reports

Figure 10: Image of Growth and Proficiency Plot from Colorado's Reporting System



These plots can be manipulated by the user to show different content areas, subgroups, or years. By allowing users to customize reports and to facilitate the presentation of a vast amount of information in a clear and simple manner, educators and other stakeholders can more easily locate findings in the data that can inform improvement initiatives.

Additionally, a comprehensive reporting system is accompanied by supporting information to help users navigate through the data and interpret findings. Innovative systems do not restrict these resources to printed reports, but take advantage of technology to produce resources such as narrated demonstrations, videos, or user guided tutorials.

Consequences and Support

As discussed with both the Advisory and Select committees, there is absolutely no intention to institute any sort of punitive consequences in reaction to accountability system results. Rather, both groups are committed to ensure that the accountability results contribute the continuous improvement process to improve the particular schools and the system as a whole. However, we are not blind to the fact that one person's support could easily be seen as another person's consequences, especially if that means restrictions on some aspects of local control. Nevertheless, both committees make these recommendations with the intention of improving Wyoming schools, particularly those performing below state expectations. The consequences and supports tied to school performance on the accountability system are multi-tiered, but the various levels are interrelated. The overall accountability level triggers a general action, but this must be further specified according the performance on the various indicators. The general actions tied to each of the overall levels are described below. The specifications of these improvement plans will need to be fleshed out with more details as the system moves towards implementation. Further, these details should be tied to the systematic efforts to improve the capacity of the schools, districts, and the state itself, described elsewhere in this report. The specific consequences and expected levels/types of support are outlined below. In Section IV of the report, we provide a more detailed description of the system of support and capacity building necessary to ensure the success of the full system.

Exemplary/Exceeding Expectations: Schools in this category should be publicly recognized and commended for their accomplishments. In order to maintain high levels of achievement and illuminate promising practices, schools in this category must file a *"communication plan"* with WDE that describes how the schools intends to document its effective practices and share these successful practices with other schools in Wyoming. This plan should be a brief document and is not intended to interfere with the school's overall success.

Satisfactory/Meeting Expectations: Schools in this category must file a *"level one improvement plan"* with WDE that is based on a close examination of the indicator scores. The level one improvement plan must be aimed at improvement goals tied to performance on the specific indicators where the school's performance was either weaker than other categories or lower than the state average performance. The level one improvement plan may include a limited number of other goals beyond the specific indicators and the plan shall include a rationale for selecting the improvement goal(s), the processes that the school will implement in order to address the goal(s), a timeline and relevant benchmarks for addressing the goal(s), and a description for how the school will evaluate its success at meeting the goal(s). WDE will appoint a liaison to monitor the school's progress at meeting the goals and to work with the school, if requested, to help support the school's efforts or to assist the school in locating additional capacity to support the school's improvement efforts. The school and district will use existing block grant funds to pay for any additional resources.

Approaching/Partially Meeting Expectations: Schools in this category must file a *"level two improvement plan"* with WDE that is based on a close examination of the indicator scores. The level two improvement plan must be aimed at improvement goals tied to performance on the specific indicators where the school's performance was either weaker than other categories or lower than the state average performance. This plan must address all areas rated unacceptable. The level two improvement plan focus only on goals related to shortcomings on the specific indicators unless there is a compelling reason to include other goals. The plan shall include a rationale for selecting the improvement goal(s), the processes that the school will implement in order to address the goal(s), a timeline and relevant benchmarks for addressing the goal(s), and a description for how the school will evaluate its success at meeting the goal(s). WDE will appoint a liaison to support the school in identifying and addressing the goals and to work with the school, if requested, to help support the school's efforts. The liaison must assist the school in locating additional capacity to support the school's improvement efforts. The district and WDE share the costs to pay for any additional resources. Schools that do not meet their improvement goals for two consecutive years under the level two plan may have their overall level changed to "priority improvement" and participate in the consequences and supports associated with that level of performance.

Priority Improvement/Not Meeting Expectations: Schools in this category must file a *"turnaround plan"* that describes how the school, along with a distinguished educator appointed by WDE and the local board of education, will radically improve its performance and must address all areas rated unacceptable. Recognizing that such significant improvement takes time (e.g., 3-5 years), the plan must specify process and performance milestones for each year that the plan is in effect. These milestones must be agreed upon by the local board of education, the distinguished educator, and the WDE liaison. The plan must identify the highest priority areas that will be the focus of the school's initial efforts, but should also discuss how the school will move beyond these highest priority indicators to other salient improvement targets. The plan shall include a rationale for selecting the improvement goal(s), the processes that the school will implement in order to address the goal(s), a timeline and relevant benchmarks for addressing the goal(s), and a description for how the school will evaluate its success at meeting the goal(s). The WDE liaison must assist the school in locating additional capacity to support the school's improvement efforts and the liaison along with the distinguished educator shall be able to direct the school and district to utilize certain improvement strategies and/or materials (e.g., curriculum). The plan must describe the resources required to carry out the improvement efforts, but must first document how existing resources will be reallocated to meet the needs described by the turnaround plan. WDE will provide the resources necessary, as authorized through this statute, to support the school's turnaround efforts. Schools that do not meet their performance improvement benchmarks under the turnaround plan for two consecutive years must hire a "school turnaround specialist" to either work with the existing school principal. Further, continued low performance may lead to termination of the principal and other staff members.

Educator Evaluation

Introduction

Like many other states, Wyoming has set out to develop a system for measuring teacher and administrator effectiveness influenced in part by student achievement. While many of the issues related to school accountability overlap with educator accountability, there are numerous specific considerations that should be addressed, which is the focus of this section. Importantly, this section is only an introduction to the very complex challenges associated with designing an educator evaluation system and does not contain the specificity needed to fully design and implement an educator evaluation system that includes measures of student academic performance. We intend for this document to provide an overview of the many issues and decisions policymakers and other stakeholders will need to consider.

Multiple Measures

While the inclusion of student achievement data (e.g. measures of student growth) constitutes a prominent element of Wyoming's initiative to reform educator evaluation systems, it should also be acknowledged that a comprehensive and defensible system incorporates multiple measures that go beyond student performance on state tests.

These may include some or all of the following:

- Direct observations of educators by principals or peers
- Student surveys
- Parent surveys

- Analysis of artifacts (e.g. student work, instructional activities, lesson plans etc.)

Such information is critical for several reasons. First, student academic performance cannot fully address all dimensions of being an effective educator. Additional information is needed to get a more complete picture of the educator's performance. Second, multiple sources of information can enhance the reliability of the outcomes. When a collection of evidence is used to make classification decisions, it mitigates the error that may be associated with any one less reliable indicator. Finally, qualitative information that provides more in-depth information about educator practices can make the results more useful and actionable. Given that a prominent claim in Wyoming's theory of action is the use of educator evaluation results to improve practice, it is important to assemble information that better allows one to understand and receive feedback on specific professional practices associated with more or less favorable academic outcomes.

Measuring Student Performance

Fundamentally, any use of student academic performance data to inform judgments of teacher effectiveness should control for prior performance. Therefore, the assessments used must produce a measure that reflects the progress or growth of the student during the period of time the teacher provided instruction. Broadly, there are two primary elements that must be in place to accomplish this goal: 1) availability of one or more suitable prior scores and 2) application of an appropriate analytic method.

To start, the structure of the assessment system should be such that one or more suitable prior scores are available. One way to accomplish this is to use a score from the end of the previous year. Given that there is an assessment at the end of each of grades 3-8 in mathematics and reading, it may be possible to use the previous year's PAWS score as a baseline for determining progress starting in grade 4. However, this assumes that the tests are highly correlated and otherwise well designed for this purpose, including content representation, breadth and depth of information, and technically defensible. We address this topic in greater depth in Section V of this document.

However, this approach is more complicated for content areas not tested annually (e.g. science and high school) or for which no suitable standardized assessment exists (e.g. physical education, art). To be sure, the 'non-tested' issue is one of the most intractable challenges facing states seeking to include student performance in educator evaluation systems. A complete treatment of this issue is beyond the scope of this document, but a summary of some alternatives more fully developed in Marion & Buckley (2011) follows[9]:

1. **Custom developed state tests**: Wyoming may elect to develop new tests to address key gaps in tested grades or content areas. An advantage of this approach is that it likely offers

[9] See: Marion, S.F. & Buckley K. (2011). Approaches and considerations for incorporating student performance results from "Non-Tested" grades and subjects into educator effectiveness determinations. Available at: www.nciea.org

maximum opportunity to create high quality assessments aligned to standards. However, the obvious disadvantage is the tremendous outlay of resources – both time and money – to develop and manage quality assessments over time.

2. **Commercially available tests**: Although some vendors offer seemingly promising standardized assessment solutions that can be flexibly administered and are less expensive that custom developed tests, this option is not without substantial risk. Most prominently, there are often serious issues with alignment and technical quality of 'off-the-shelf' tests.

3. **School/ teacher created tests:** Allowing schools or classes to develop assessments could serve as a professional development tool for educators and should promote alignment between instruction and content. Another advantage is the potential to measure more complex knowledge and skills than a selected response tests. However, both the quality and the comparability of tests developed at the school or class level is a very significant issue. Also, this approach may be more corruptible than other approaches.

4. **School-wide attribution:** In the absence of current and/or prior test data for the class/course of interest, it is possible to assign a school rating to the teachers with missing data. This alternative does introduce serious concerns that linkages between services and outcomes may be less direct. However, others have argued that such an approach increases engagement and cooperation of personnel throughout the school.

5. **Student learning objectives**: Some states are considering student learning objectives (also termed student growth objectives.) Broadly, this approach involves teachers drawing on classroom-based information to establish goals for individual students or the class. The teacher then evaluates student and/or class progress toward these outcomes. This approach is appealing in that it has great potential for both educator and student development through the process of establishing and pursue meaningful learning outcomes. However, comparability and corruptibility are non-trivial threats to guard against.

Inclusion and Attribution

Attribution refers to an essential claim in the theory of action that educator practices influence the academic performance of students. To address this, Wyoming must be able to link student outcomes to educators and assemble evidence that demonstrates a credible connection between these elements.

Teacher/Leader of Record

Addressing attribution starts with determining which teacher/leader should be held accountable for a student's performance. This is often referred to as *defining the teacher of record*. A suitable definition - and an accompanying data system that permits operationalization of this definition - should establish the conditions and circumstances governing the connection of educators with classes and account for the variety of learning environments in Wyoming schools.

For example, the Data Quality Campaign (DQC) (2010a) advises states seeking to use assessment data to inform educator evaluation to:

- Account for contributions of multiple educators in a single course
- Enable teachers to review rosters for accuracy
- Account for schedule changes and variable class environments such as virtual classes or labs
- Link attendance records with teachers to track actual days of instruction

Using a modified version of the high-level ‘framework’ for defining teacher of record offered by the DQC (2010b) a sample operational definition for Wyoming might include the following:

- The educator/ leader roles included (e.g. certified educators, academic coaches, mentors etc.)
- The amount of instructional time to establish a link (e.g. responsible for at least 50% or more of instructional time)
- Courses/ environments covered (e.g. courses for which there is an associated, valid test score)
- Prior measures required (e.g. at least two prior valid PAWS scores in the same content area).
- Other conditions (e.g. continuous enrollment requirement)

<u>Missing/ Incomplete Data</u>

Another ‘data issue’ to address is missing and/or incomplete data. This situation exists when any of the following occur:

- One or more prior (pre) test scores are missing
- The current year (post) test score is missing
- The student is not continuously enrolled in a single building/class throughout the term of instruction
- The student record is missing or incomplete (e.g. test scores but no identifier)

Missing data can impact the precision and stability of the growth analysis and introduce systematic bias in the resulting estimates (Braun et al, 2010). Moreover, it is generally acknowledged that data are not Missing At Random (MAR), meaning that it is likely that the performance of students with missing or incomplete data differs systematically from those with complete records. Consider, for example, that mobility rates are typically higher for economically disadvantaged students compared to other students.

When all or part of a record is missing, there are a number of potential methods to address this. One solution is to simply omit the records. This approach may be simple to understand and straightforward to implement, however, it is likely most vulnerable to potential introduction of bias for the reasons noted above. Alternately, Wyoming may implement one of several approaches to data imputation – or using a statistical method to populate the missing value(s). Imputation methods range from simple (e.g. replacing the missing value with the mean value of all existing data) to more complex (e.g. using an algorithm to predict the likely value of the missing value based on patterns in the existing data).

There is no single or best approach to dealing with missing data. In general, we recommend Wyoming consider the following steps to address this threat moving forward.

- Identify business rules informed by impact analyses that clearly define what data are usable and which are not. Consider issues such as:
    - What is the minimum group size to calculate a class/school growth estimate?
    - Regardless of group size, what is the minimum inclusion rate to calculate an estimate? Inclusion rate refers to the proportion of students in a class or school that 'count' in the analysis. For example, if only 10 students in a class of 30 are included, this may meet the n-size rule, but may not be judged sufficient to represent the overall class effect.
    - How long must the student be enrolled in the class to 'count' in the computation?
- Investigate the extent that data are missing for districts, schools, and classes. Seek to understand patterns of missing data for various levels of performance and by subgroup. Such analyses will help determine the extent to which data are MAR or differ in a systematic manner.

Multiple Educators

As mentioned earlier, another issue to consider is how to handle circumstances where students receive instruction from multiple educators. There are three general cases that lead to this occurrence. First, the student may receive planned, ongoing instruction from multiple teachers, as with a team teaching approach or scheduled support sessions. Second, changes can occur throughout the year, such as a leave of absence for the primary instructor or the student transitions to another class. Finally, additional instruction can occur in a variety of contexts, such as when a student receives tutoring outside of class. Whatever the case, multiple sources of instruction will likely have an impact on student achievement.

Some researchers have hypothesized that a 'dosage' model may be appropriate in such circumstances. That is, if Ms. Smith provides 70% of instruction and Mr. Jones provides 30% of instruction, then the outcomes are assigned to the educators consistent with the proportion of instruction provided. While it may be useful to research the feasibility of this approach, we are skeptical that proportional contribution to instruction can be captured with precision, particularly when it is unscheduled. Also, it will be necessary to create potentially complex connections in the state data system to account for this. It is important to consider that the proportional contribution to instruction may not be governed by time alone. For example, an hour spent introducing new concepts to a class may not represent the same 'instructional contribution' as an hour spent overseeing time allotted for student directed study. Finally, the research on attributing a student's academic performance to teachers and leaders is emerging – even for the least ambiguous circumstances when the teacher of record is well defined. Much less is known about the credibility of results based on proportional attribution of scores.

We advise Wyoming to proceed with caution in exploring a 'dosage' model, ensuring the information is suitably trustworthy and the results are scrutinized carefully, particularly with respect to evidence of reliability and validity presented later in this document.

Causal Attribution

As stated previously, the use of student performance data to inform evaluations of educator effectiveness assumes at least a partial causal link between teacher performance and student outcomes. Establishing such links are problematic in light of research which suggests that though teacher influence on student learning is significant and persists across years, isolating that contribution using large scale assessment using observational data is difficult, if not impossible to accomplish. Numerous published writings by scholars on the subject over the past decade support this (see, for example, Raudenbush (2004); Rubin, Stuart, & Zanutto (2004); Linn (2008); Rothstein, 2009; 2010; Betebenner & Linn (2010); Briggs & Domingue (2011)).

In light of this, the use of student growth as a component of a high-stakes evaluation model demands additional evidence to validate a claim of effectiveness with regard to instruction. The collection of such evidence will help to bolster the credibility of the model and validity of the outcomes. Validation of effectiveness claims is a non-trivial task and typically involves engaging in systematic data collection and research to both strengthen the association between the hypothesized antecedent (i.e. quality instruction) and the consequent (i.e. increased test scores) and to rule out rival explanations for the outcome.

A good starting place for a program of research is to seek to determine a 'proof of concept.' That is, in the best case with at least a group of 'consensus quality educators,' (necessarily defined by judgment and existing criteria) what is the impact on student achievement? To what degree does this differ by content area, for students of various ability levels, among special populations, and over years?

Attribution claims can be further strengthened by addressing the sensitivity and bias of model results. For example, in their review of analyses of educator data in the *Los Angeles Unified School District,* Briggs and Domingue examine the extent to which the original model may have been misspecified, by investigating whether a student's teacher in the future could have an effect on the student's prior test performance (2011). Naturally, a strong 'reverse association' erodes confidence that the model is well suited to support claims. Briggs and Domingue also introduce variations in model specifications to explore consistency of ratings and examine outcomes with respect to confidence intervals to evaluate the precision of the estimates and the basis to claim the resulting classification is accurate. These analyses provide examples of the types of investigations that can serve as components of an overarching research agenda to explore the credibility of causal claims.

Reporting Outcomes of Educator Evaluation Determinations

Another critical decision for the educator accountability system will be to define the type and manner of reported results. This starts with clearly establishing the performance levels that must be produced and the purposes for which they will be used. In general, there is a tension between reporting high-level results that are more reliable and the desire to report more nuanced but less precise outcomes for multiple indicators. For example, there will be a much higher level of confidence in classifications of class effects as low, typical, or high compared to a class effects described on a ten point scale from 1 (ineffective) to 10 (highly effective). In the latter case, stakeholders may regard this information as useful to understand more fine grained degrees of difference, but such a scale may carry only the appearance of precision that is not supported by evidence, particularly for adjacent ratings.

The same issue is generally true for reporting units. That is, results for individual content areas or classes will be much less defensible (and results based on strands or subscores will be almost certainly indefensible) than aggregate results for multiple classes. The goal, of course, is to find the balance between the necessary specificity of outcomes and an acceptable level of precision. As a matter of best practice, is advisable to privilege technical defensibility, in order to provide the best case for results to be meaningfully interpreted and utilized.

Finally, it is important to consider how to combine indicators and set performance thresholds. Once the key elements that will influence evaluations are identified and decisions are made about the 'weight' of each component, it is possible to combine the indicators in a manner similar to the alternatives described in the design section of this document.

These decisions are closely connected to the consequences and rewards that are indentified. In general, the higher the stakes, the higher the standard of evidence should be regarding the classification accuracy of the system. For example, it may be appropriate to require multiple years of low ratings to support a high-stakes decision such as termination or reassignment.

Sources of Error

There are multiple sources of error that may impact the precision and, consequently, the usefulness of model result[10]. The first is measurement error. Measurement error refers to the extent to which individual assessments in the evaluation system produce stable and consistent results.

Another threat is related to sampling error. This refers to variations in the population at the unit for which inferences will be based – the district, school or class. Sampling error is known to promote substantial fluctuations in school scores that can be unrelated to actual school performance (see e.g. Hill and DePascale, 2002) and it has the potential to introduce a great deal more uncertainty in class outcomes. This is particularly relevant given that students are rarely, if ever, randomly assigned to teachers. Sampling error is directly related to the number of observations - as the sample size increases, the variability reduces. Therefore, the problem is somewhat assuaged when computing a growth score for a school across several teachers and grades.

Yet another potential source of error is related to model specifications. Researchers have found that estimated effects are sensitive to model assumptions and specifications (see e.g. McCaffrey et al, 2003). In other words, adjustments to model characteristics, such as adding, deleting, or differently defining variables, will very likely produce dissimilar results.

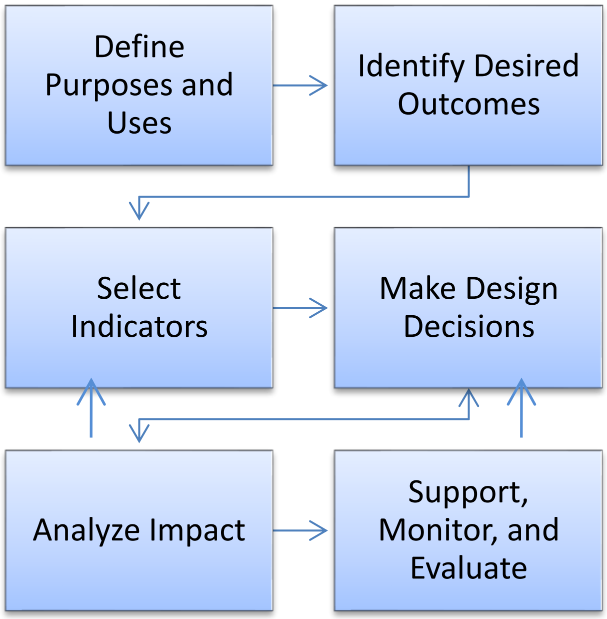
[10] Information regarding sources of error and threats to utility addressed in Domaleski and Hill, 2010.

Implementation Plan

The design and implementation of a reliable and valid system to evaluate educators involves addressing many complex challenges. In this section, we summarize the most important steps in the process to provide the basis of an implementation plan for the state.

Figure 11 shows the six major steps involved in implementing the educator evaluation system. This process is not linear. We recommend a process in which impact analyses and ongoing evaluation are used to gauge the adequacy of the model and inform appropriate changes to the indicators used in the model or refinements to the design.

Figure 11: Key Components in Design and Implementation of Educator Evaluation System



The first step in the process is to clearly define the purpose and uses of the system. As detailed earlier in this document, the intended goals should be reflected in an explicit and credible theory of action that makes clear the assumptions about what the state hopes to accomplish with the educator evaluation system and how the process will promote the desired outcomes, including the mechanisms that are hypothesized to promote these goals. This shapes subsequent decisions such as what information to include, how to report outcomes, and how to set performance expectations. This also helps clarify if/why certain requirements are important, which helps prioritize the elements that are most central to the success of the initiative.

Next, it is important to clearly define the desired outcomes. This includes identifying what information will be produced and how it will be communicated to all stakeholders. For example, will the system produce performance classifications? How many levels of classifications will be produced and what will they mean? For instance, if the top classification is intended to qualify an educator for merit pay or the bottom classification leads to termination, this must be clarified from the start in order to better understand what information is needed and how performance standards should be established. Additionally, what content areas will be covered? Will the classifications combine content areas for each teacher or be specific to each content area? In what area should educators, leaders, parents, etc. receive feedback from the system (e.g. academic growth of students, professional practices of educators etc.)? Only by laying bare all the intended outcomes of the system and the target 'audience' for each can developers ensure design decisions are made that support these outcomes.

Next, the state should identify the indicators that are central to supporting the goals and outcomes of the system. This is likely to include the academic performance of students, which, as noted previously, is much more difficult to address in content areas where a series of technically defensible, standardized, summative assessments are not administered. It may also include qualitative measures of educator performance such as observations of instructional practices, peer ratings, or surveys. In each case, it is important to determine what information is needed to support the claims and uses, whether this information can be obtained in a manner that is not overly burdensome to schools and systems, and if this information is likely to be sufficiently credible to support the intended claims.

Once the potential indicators are selected the next step is to determine how the information will be used to produce outcomes. This involves selecting a growth model and resolving the requisite specifications and decisions about the model (see growth section in this document for more detailed treatment of this topic.) It also involves determining how much weight or influence will be given to certain indicators (e.g. will qualitative evaluations count more, less, or about the same as academic performance indicators?). Whether to combine indicators both within and across categories is also resolved in the design phase. Yet another prominent design decision involves setting performance standards. This defines the minimum expectation for adequate performance or how well an educator must perform to attain designations intended to reward exemplary performance or that signal performance that is below standards. Finally, in the design phase it is important to identify mechanisms that are likely to bolster the reliability and validity of outcomes. For example, using results that reflect an average over multiple years may be regarded as preferable to a single year to enhance reliability of results. Or, it may be necessary to adopt different rules or procedures for educators in certain schools or class environments, such as those teaching in alternative schools.

In order to make the best decisions about the suitability of the model, including identification of trustworthy indicators and appropriate design decisions, it is critical to engage in ongoing data analysis. These analyses should include a review of the distribution of outcomes for all proposed reporting units and aggregated to various summary levels. Special attention should be given to examining results based on differences in student populations (e.g. are results different for educators in schools serving a high percentage of impoverished students?) and based on differences in indicators (e.g. are results substantially different for selected grade or content areas?). All indicators should be carefully piloted and results should be investigated for reasonableness and compared to any credible existing information to assess the validity of outcomes. For example, if a pilot of peer surveys or a trial of proposed observations of instructional effectiveness reveals little variation in outcomes (i.e. all or nearly all teachers are rated effective) then the credibility of the indicator is called into question. This may necessitate removing or changing indicators, reweighting model components, and/or adjusting performance expectations.

Finally, a comprehensive implementation plan should include a process for ongoing monitoring, evaluation, and support. This includes but goes beyond producing impact analyses as described in the previous stage. In addition to examining year to year changes in outcomes, the evaluation plan should investigate the claims and assumptions in the theory of action. For example, are educators and leaders using the information to improve practice? Are rewards effective incentives? Are remediation and support strategies effective in improving outcomes? A systematic process to collect evidence and evaluate model claims will help state leaders identify refinements to the model to improve effectiveness.

Student Accountability Considerations

Introduction

Senate File 70 directs the State Board of Education to review an alternative to the current body of evidence system with a goal of replacing the current BOE system for school year 2012-2013. The legislature directed the SBE to consider using end-of-course (EOC) tests that could be used as an alternative to the Body of Evidence (BOE). Since the BOE is a student accountability system designed to determine if students are eligible for high school graduation, we assume that the EOC tests are expected to support graduation decisions. In this section we discuss some considerations for creating/modifying a high school graduation system.

First, we note that it is beyond the scope of this report to make specific recommendations about a student accountability system since this issue was addressed only peripherally by the Select and Advisory Committees. Rather, we outline the steps necessary for creating an EOC-based student accountability system and highlight key considerations for the Select Committee and other stakeholders. In the course of revising the current graduation system, subsequent legislation should define a process for making critical decisions about the various components of such a system. This legislation should explicitly articulate the degree to which the new legislation is replacing or working within the context of existing Wyoming graduation statutes (W.S. 21-2-304 and the State Board Chapter 31 Rules). This process should undoubtedly include key stakeholders, as part of a design committee, such as local school board members, district and school leaders, teachers, guidance counselors, businesspeople, higher education representatives, and students. These stakeholders should be guided through a process where they can wrestle with the following key components of developing a student graduation accountability system:

- Definition of a Wyoming graduate
- Knowledge, skills, and dispositions
- Accountability Decisions
- Assessment system
- Support and Interventions

What is a Wyoming Graduate?

The most critical aspect of developing a student graduation accountability system is to define a Wyoming high school graduate. The design committee should spend appropriate time developing this description and likely should solicit significant input prior to moving forward. Ideally, the goal is to develop a shared understanding of what it means to be a Wyoming high school graduate.

The next step in the process is to describe the knowledge, skills, and dispositions (perhaps) that further specify the definition of a Wyoming graduate. These are often the high school content standards in the various subject areas. But if things like dispositions (e.g., persistence) are included, the design committee should specify these non-content areas such that students, parents, and teachers are clear for what students are being held accountable. The design committee should also wrestle very important considerations such as how well the students need to perform on the standards in order to graduate and whether or not students should have to perform up to these expectations on all standards or content areas, a targeted set (core) of content areas, or some combination of these two possibilities. Of course, there are other possibilities that must be determined by this committee.

A Process for Thinking About Student Accountability

While SF 70 recommends that the State Board consider implementing an EOC system to potentially replace the current BOE system, it was silent on many important details. Before designing an EOC assessment system, the design committee, State Board, and perhaps the legislature will need to define the accountability rules of the graduation system. This follows naturally from the discussion of the required/expected knowledge, skills, and dispositions. There are many such accountability and assessment decisions to be made, including:

- What framework or approach will be used to organize the EOC exams?
- Which courses will include a state EOC exam?
- Which standards will the tests be designed to measure?
- What are the participation rules for any or all of the exams?
- At what level should the passing scores be set?
- What consequences will be associated with the results?
- Will retesting be included? How many opportunities?
- Can other sources of evidence replace EOC scores?
- Will there be an appeal process for students not meeting graduation standards on the testing system?
- How will the system address issues of student mobility?

The SBE and the design committee will first need to define a framework for organizing the EOC exams[11]. It is doubtful that the legislature intended to authorize creating EOC exams in every possible high school course. W.S. 21-2-304 and Chapter 31 required that students meet

[11] For more information on state practices and alternatives relative to using EOC tests in accountability see: Domaleski, C.S. (2011). *State End of Course Tests: A Policy Brief.* Paper commissioned by the Council of Chief State School Officers Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards.

standards in all nine content areas included in the “basket of goods.” Even though this narrows the range of possibilities from all possible courses to an exam or set of exams in each of nine content areas, this will still be a significant expense and will require considerable resources within WDE and LEAs to successfully implement such a system. Therefore, we are interpreting SF 70 to mean that the EOC should focus on key courses within the four core subject areas of mathematics, science, social students, and English language arts.

With guidance from the Select Committee, the design committee will need to identify the courses for which EOC exams will be created. However, existing rules require that students demonstrate proficiency in five of the nine content areas. Therefore, existing statute and rules will need to be amended or these EOC exams will have to fit within the existing Chapter 31 framework. For example, all or some of the EOC exams could be required components in each district’s Body of Evidence system. Having such a framework will help with decisions about whether students will be required to pass any or all of these exams in order to graduate.

The current Wyoming content standards, as well as the Common Core State Standards (CCSS), are domain-based and not tied to specific courses. However, in order develop EOC exams, it will be important to identify the eligible content and skills for each of the exams. This might mean simply identifying existing current standards that will be tested in each of the courses or developing content frameworks specific to each course. In either case, it will be critical to the validity of the exams and to the transparency of the system for the State to explicitly identify the eligible knowledge and skills for each of the exams.

Once decisions are made about the courses for which the EOC exams will be developed and what they will measure, the design committee must determine the participation rules for the various exams. For example, will all students be required to take all courses for which there is an EOC exam and/or will all students enrolled in an EOC course be required to participate in the exam?

In addition to expected consequences associated with these exams (i.e., they will count towards graduation decisions), there are other decisions to be made related to consequences. For example, will students be expected to pass the exams in order to pass the course, will the exams be required to carry a specific weight in the course grade, or will the decisions about how the EOC exams will factor into course grades be left up to locals? Any decision to count the EOC exams as any part of the course grade will have important implications for the timing of the testing window and required turnaround time for scoring. As noted above, situating these decisions within a larger framework (e.g., BOE) will lead to more coherent policy.

If there are consequences associated with individual exams (e.g., passing the course or if students are expected to pass a specific set of exams to graduate), the design committee and policy makers must deal with the issue of retesting. Essentially all states that use exam-based approaches to graduation decisions permit at least one, and often many, retest attempts. If the exams are to count in course grades, this raises many tricky logistical and fairness issues. But even if the exams are not included in course grades, the issue of retesting can be much more challenging when dealing with EOC exams compared to a more common end-of-high school exam[12].

The issues of alternate sources of evidence and potential appeal processes are somewhat related to the retesting issue. The design committee and policy leaders will have to determine if other sources of evidence (e.g., portfolios or projects) can substitute for any or all EOC exams. If so, a design committee and policy leaders would have to decide if such alternatives should be available to all students or just certain groups of students (e.g., special education, ELL, migrant). Additionally, it will be prudent to plan for an appeal process for students who do not meet graduation requirements. Related to the alternate evidence issue, the design and policy committees will need to decide how to handle students who move into Wyoming after any or all of these exams are typically offered. A likely approach will be to use the student's transcript to provide "alternate" evidence that the student met or did not meet the graduation evidence represented by specific EOC exams. Again, it makes sense to address these major policy issues within a larger graduation requirement framework.

Relationship to the Full Assessment System

Current practice in Wyoming involves administering one assessment in high school for each reading, mathematics, science, and writing. If an EOC testing system was implemented, it would make little sense to continue to administer the end of domain tests as is current practice, but to rely on the EOC system to serve as the school and educator accountability assessments for high school. The Advisory committee should study and make recommendations about how best to use the EOC tests in the school and educator accountability systems.

A Process Note

As illustrated above, there are many thorny issues to address in the design of a student accountability system. To that end, we recommend that the current Advisory Committee, along with perhaps some additional ad hoc members, be invited to serve as the basis for a design committee that reports to the State Board of Education.

[12] Note: We are definitely not recommending a single set of end-of-high school exams, but just pointing out the contrast.

SECTION IV: SUPPORT, CAPACITY BUILDING, AND CONSEQUENCES

Support, Interventions, and Capacity Building

The Advisory Committee recognizes that an accountability system is only valuable if it leads to, or at least facilitates improvement in both student and school results. The accountability system itself cannot improve student and school achievement, but it should be designed to both incentivize the "right" behaviors and provide results that are specific and informative enough such that school leaders and other stakeholders can learn about the educational aspects under their control that might need improvement. One of the things we know well from educational psychology is that task-specific feedback is more likely to lead to improved performance than general feedback. We have no reason to believe that organizations would act differently than individuals in terms of the response to specific or general feedback. This section of the accountability framework, based on extensive discussions and input from the Advisory committee, describes some of the supports and interventions necessary to realize the type of improvement and level of achievement envisioned by Wyoming policy makers. While it is tempting to collapse all supports and interventions into a single topic, the Advisory Committee recognizes that it is important to address each of the following levels in a comprehensive support system:

- Support/intervention for low performing students
- Support/mentoring for teachers needing to improve
    - Induction for new teachers and leaders
- Support/mentoring for school leaders
- Capacity building for schools and districts with lower than acceptable levels of achievement or growth
- Capacity building for the state as a whole to support continuous improvement
- The role of institutions of higher education in building capacity and preparation especially in terms of P-16 coordination

Further, the Advisory Committee recognizes that several aspects of such a support/capacity building system are already provided for in the school funding formula. The committee, however, strongly suggests that these aspects of support/improvement be addressed comprehensively along with the development of an accountability framework.

Elmore is quite eloquent and persuasive in outlining at least one aspect of the challenges we face. While there is little talk of "stakes" in the sense of what we commonly think of as high stakes (e.g., takeovers, firing school leaders), the labels placed on schools via the reporting of accountability system results and the public dissemination of such results are seen as stakes by many in the system. Our charge is becoming clearer. We must insist on a system that allows schools to develop the capacity they need to affect the instructional core. Just as we have argued for formative assessment to help students know where they stand relative to key standards, we also need tools to assess the capacity of schools to enact key reforms and interventions. Elmore (2004) reminds us of the challenge we face in our work:

> *Hence, stakes work, if they work at all, by mobilizing and expanding capacities in high-capacity schools and creating potential demand for capacities outside the organization in low-capacity schools. In the latter case, if there are no capacities to bring to the organization, there is little reason to expect the organization to do*

> *anything other than to make incremental adjustments to already unsuccessful practices (p. 289).*

In this 2004 chapter, Elmore goes on to outline five principles of accountability system design. While all of the principles are worth considering, the fifth principle is especially pertinent to the work of the Advisory Committee.

> *The fifth principle is the reciprocity of accountability and capacity—for each increment in performance I require of you, I have an equal and reciprocal responsibility to provide you with the capacity to produce that kind of performance (p. 294).*

It is important to think of this as a multi-level challenge. For example, the "I" could be the teacher and the "you" could be their student(s). Similarly, the "I" could be the principal and the "you" could be the teachers, and so on. The point is clear. Each level of the system that is imposing any sort of accountability on the level below is responsible for providing the capacity for that level to succeed.

Building Capacity in Wyoming Schools

Given this framework for thinking about accountability and capacity, we discuss the multiple levels of capacity needs, starting from the students and working up to the state level. This is not the place to present a definitive plan for capacity building at all levels of the accountability system. Rather, our goal here is to outline the key considerations for each of the levels and to argue that the State convene appropriate advisory groups and relevant agency personnel to develop detailed plans (including cost ramifications) for addressing these issues in the context of a comprehensive accountability system.

Capacity Building for Schools and Districts

Given that accountability system is focused first at the school level, improving the capacity of schools will require considerable effort and support. The accountability system itself must be designed to incentivize appropriate activities, but as importantly, must yield information that school leaders and educators can understand and use to help identify areas in need of improvement. As with students, information must be specific to the particular initiative and focal area. Information should also be presented for current and multiple years to avoid having schools act on what might not be reliable yearly information.

Further, the accountability system should be focused on the highest leverage indicators, in terms of bringing about significant improvement in the rates of college and career readiness. This does not preclude the reporting system from including a broad array of process and outcome indicators. However, the accountability system should help schools develop a clear focus on those indicators deemed to be most important. This would send a clear message to schools about what is most valued and what levels of performance are deemed acceptable. If designed well, the reporting system should allow the schools and perhaps capacity building personnel to use this additional information to help improve performance on the accountability indicators. In other words, the information included in the reporting system should be linked through a theory of action to the accountability indicators. For example, we discussed holding schools accountable for graduation rates, but including credit accumulation at the end of 9th grade in the reporting system because of its clear link to the accountability indicator.

We must ask in terms of capacity building about additional support and capacity building needs required by schools beyond those targeted for teachers and school leaders. It can be argued that schools are simply collections of individuals, so that if we focus on students, teachers, and principals, is there any need to worry about building "school capacity?" We argue that just like if both spouses in a marriage pursue counseling as individuals, there is generally still a need to pursue marriage counseling to address "system" issues. Similarly, we argue that the system issues of a school should be addressed as well.

The capacity building needs for schools could be considerable and highly varied. Therefore, an effective set of supports organized at the district, regional, and/or state level should be able to differentially respond to the varied needs of schools. This suggests a more nuanced approach than simply having all schools follow the same school improvement steps. A regional approach that included some sort of intermediate level service agency with enough capacity to adjust to the varied and multiple needs of schools could be one approach for increasing school capacity in WY. However, many states have such agencies (e.g., BOCES) and it would be worth careful study of these intermediate agencies in other states to identify the most effective organizational and educational strategies before adopting such an approach in Wyoming.

Schools have specific cultures and high functioning schools have cultures where data are used to identify goals, design interventions and strategies, create or select tools for monitoring the progress toward goals, evaluate the success at meeting the goals and then starting the cycle again. This problem identification, hypothesis testing, and evaluation is the work of high performing schools and the work that we hope becomes enculturated in all schools. The advantage of this hypothesis-testing approach is that it quickly moves away from a central focus on a one size fits all solution, but helps to build the capacity so that schools and districts develop the tools and techniques to address a range of problems that might be faced by schools.

Support/intervention for low performing students

Students will perform "poorly" for a variety of reasons and conditions. The first step is to be clear about what we mean by "poor performance." Even taking a simplistic definition of poor performance, such as scoring below proficient on PAWS, leads us quickly into a myriad of possible diagnostic and intervention paths. While the information available from a summative assessment is necessarily limited in terms of student diagnosis, the inclusion of student longitudinal growth results can allow the school to determine if the student is low achieving, growing slowly (relative to peers), or both. However, schools will need considerably more fine-grained information to be able to better understand students' strengths and weaknesses if they want to implement systematic approaches for improving student learning. First, schools should not be waiting until the summative assessment results are returned at the end of the school year or in the summer to find out that students are performing below expectations. Additionally, it is highly unlikely that a two or three times per year "benchmark adaptive assessment" will provide specific and frequent enough information for diagnosing and monitoring student achievement. Schools will need to implement systematic approaches for helping students improve their performance, including (but not limited to):

- Appropriate support and interventions for special education and English language learners,

- Formative and classroom assessment tools useful for ongoing progress monitoring and interventions,
- Employing a Response-to-Interventions (RTI) or similarly systematic approach for diagnosis, intervention, and monitoring,
- Differentiated instruction within classrooms and additional support services outside of classrooms for targeted instructional areas, and
- Creating "extra time" opportunities such as after school and summer school enrichment opportunities.

Any of these approaches should work to encourage the development of student agency and metacognitive strategies so that students develop internal capacity to learn to help themselves. Of course, a discussion of supports for students leads quickly to the recognition that, as Elmore noted, high-performing schools already have the capacity to address many or all of these examples of student supports, whereas low-capacity schools do not. This can really be viewed as a problem solving or hypothesis testing enterprise in that the first step is to figure out the problems, pose strategies, and find the capacity to address the problems. This really should be seen as the work of schools and not as extra work.

Support/Mentoring for Teachers Needing to Improve

Many of the approaches highlighted above for students assume that a high quality teacher will be in place to provide such services. As Elmore indicated, unless something shifts in the instructional core, student learning is unlikely to improve. Teachers have the major responsibility for improving the quality of the core, but many need help to enact the high quality instruction needed to bring about high levels of student learning. This is especially critical if expectations are to be increased such that all students leave high school ready for college or careers. Further, if Wyoming adopts the Common Core State Standards (CCSS), the need to raise curricular and instructional expectations will be immediately apparent. The accountability system must then provide information that is specific enough to enable schools and teachers identify strengths and weaknesses for targeting improvement efforts.

The Wyoming school funding model currently includes provisions for an instructional coach at each (or most) buildings. This is certainly a good start towards building increased capacity among Wyoming's teachers. Further, while having such a resource in each building would be considered a luxury in many states, it will likely not be enough to raise performance to levels heretofore not seen in most states. It has been well documented that many or even most teachers lack the content and pedagogical content knowledge to engage students in tasks that require students to wrestle with complex subject matter. There will need to be considerable training and support to help Wyoming teachers fully understand the curricular and instructional ramifications of the CCSS. While much of this support must happen locally, it would make considerable sense to capitalize on collective resources and expertise to help meet this enormous need.

In addition to the professional development work that must occur for existing teachers, there needs to be a considerable improvement in the quality of new educators coming out of teacher education programs. Once these new teachers enter the workforce, schools and districts need to support the continued development of these novice teachers with high quality mentoring and induction systems for new teachers and leaders.

Support/Mentoring for School Leaders

Most school reform leaders argue that a school leader is the linchpin of educational improvement. While it is possible for schools to be somewhat effective with a less-than-effective leader, it is almost impossible for a school to be effective with an ineffective leader. Similarly, an effective leader does not guarantee an effective school, but it certainly improves its chances. To hammer this point further, KIPP Schools, the highly successful charter organization, will not open a new school unless it has a well-trained principal to lead that school. Unfortunately, public schools do not have the luxury of waiting to open schools until a high quality principal is in place. This heightens the need to ensure that current principals receive the training and support they need to become highly effective instructional leaders and to improve the pre-service training provided to principal candidates before they can lead schools.

There is a pressing need to improve the capacity of school leaders in Wyoming and in most other states. Unfortunately, there are few, high quality opportunities in Wyoming to improve the capacity of current and future schools leaders. The isolated nature of schooling in Wyoming does not help the situation. Wyoming's John P. Ellbogen Leadership and Advocacy Institute is one notable professional learning opportunity for current and future school leaders, but it is not enough. A much more systematic approach will be required to recruit, train, mentor, and support current and future school leaders. In particular, district superintendents need training on how best to identify, train, and mentor new leaders.

There are several models on which to draw, led by the work of the Wallace Foundation, Interstate School Leaders Licensure Consortium (ISLLIC), and others, but it will be important to design a school leadership training and support network tailored to Wyoming's context. Additionally, the University of Wyoming will need to be engaged as the primary institution for providing pre-service education to prospective school leaders.

Capacity Building for the State as a Whole to Support Continuous Improvement

It is one thing to consider support and capacity building for individual schools, but given the goals associated with the proposed accountability system, the capacity building to meet these accountability goals will require a different form of support never seen before. Therefore, it does not make sense to operate in a reactive mode whereby the State or other provider tries to rush around the state putting out "spot fires." Rather, the state-level approach should be much more proactive by identifying the highest-leverage and highest-need topics on which to target the capacity building at the state level. If we think about the state as a system, systems can get smarter by aggregating the knowledge gained. Somehow, the knowledge gained from working at both the micro (school/classroom) and macro (region/state) levels needs to be aggregated and shared so that all in the state may benefit. One way to think about a reformed capacity building approach is to take seriously Elmore's 4th and 6th principles of the Instructional Core:

> ***Principle #4:*** *Task predicts performance.*
>
> ***Principle #6:*** *We learn to do the work by doing the work. Not by telling other people to do the work, not by having done the work at some time in the past, and not by hiring.*

In the case of building statewide educational improvement capacity, we should think of "task" more broadly than intended by Elmore's original formulation as an instructional or assessment tasks that leads students into profound interactions with meaningful content and skills. However, we do not need to stray from this formulation too far. The tasks could be those sorts of activities and products that bring teachers and leaders into "profound interactions" with meaningful school improvement activities such as using data to inform decisions or creating strategies for improving the quality and rigor of mathematics instruction. Elmore's sixth principle, we learn by doing, applies to adults as well as to students (perhaps even more so!). Therefore, the state needs to structure professional learning opportunities that are far removed from the typical "sit and get" professional development sessions.

One approach, that could be done regionally or at the state level, would involve creating networks of schools and districts interested in working on a particular issue or challenge. The Body of Evidence (BOE) Activities Consortium serves as one stellar example of a network of districts that came together to produce an important set of products, but more importantly, to increase the learning of the participants by doing the work! For those who do not remember or have come to the state more recently, the BOE consortium was a network that at its peak included essentially all districts as fully participating members, had full support (and leadership during the early years) from WDE, and high quality expert consultants. All three pieces were critical to the success, and we should be mindful that all pieces either need to be in place in the formation of any new networks or we should have a clear and defensible rationale for suggesting modifications to this approach. This is not to say that there is any single approach that will work well to improve system capacity, but we should be thoughtful in what is suggested, especially in terms of any apparent shortcuts. On the other hand, lest we appear too parochial, we would be wise to recognize some great success in capacity building examples from other states and countries. Massachusetts is one state that comes to mind from which we might find some good examples, but the experiences from Queensland, Australia, Ontario, Canada, and Finland all bear examining.

The main point of this discussion is to emphasize that the State needs to design and implement a well-conceived strategy in order to significantly raise the levels of achievement across the state. This strategy must be comprehensive and have the resources (especially in terms of expert leadership and support) allocated to support and sustain these initiatives. The Advisory Committee recommends that the State require and support a capacity-building advisory task force to help design a structure (or structures) for significantly increasing the capacity among educational personnel, institutions, and ultimately students in the state of Wyoming.

A large scale initiative to support educational improvement in Wyoming must follow a well-articulated theory of action, which the Advisory Committee recommends building around at least the following four principles:

1. Increasing student learning and minimizing achievement gaps is primarily a function of increased engagement and improving instructional quality.
2. Instructional quality depends on increasing the expertise of teachers and leaders which requires more open, public practice.
3. Instructional quality can only be improved when leaders know what quality instruction looks like and use that knowledge to support ongoing improvements e.g. the fundamental purpose of educational leadership is the improvement of instruction with everything else being instrumental to that purpose.

4. All system practices and structures must support the ongoing improvement of instruction which must be a focus of the system (Fink and Markholt, 2011).

The Advisory Committee has discussed some potential approaches to address these challenges internally, but in order to move forward, more input from a variety of Wyoming educational stakeholders is required. The primary goal would be to link and connect the various entities in the state to develop a systemic, statewide capacity development plan including the University, WDE, districts and other organizations who have an interest in this issue. Therefore, we suggest the following steps to address the capacity needs in Wyoming:

1. Create and convene a broad-based Technical Advisory committee of stakeholders, including members of the Advisory Committee, and outside experts who have done this type of work in other countries, states, and organizations. The point here is that it takes expertise to create expertise, so we need experts who have done this at large scale.
2. Conduct a comprehensive needs assessment by surveying district and school leaders, teachers, parents and others. Additionally, analyses of existing school and state performance data as well as other forms of documentation should be examined and contribute to this needs assessment. This would help identify specific problems of practice to focus on especially in the areas required by the accountability legislation.
3. Conduct a survey of existing capacity within Wyoming and then conduct a "gap analysis" to determine the gap between current needs and current capacity.
4. Use the Technical Advisory Committee to review the needs and capacity assessments and design approaches to build the necessary capacity across the state.
5. If the decision is that there is not enough internal capacity within WY to solve these challenges, determine the best way to secure external capacity-building support within Wyoming, such as through the RFP process or some other approach.
6. Report to the Select Committee during the next interim to determine the best way to significantly improve the educational capacity of Wyoming's leaders and teachers.

The Advisory Committee strongly supports the calls for enhanced accountability. However, as noted elsewhere in this report, we recognize that accountability alone will not result in the increased student learning desired by the legislature. The accountability indicators will certainly signal the desired levels of performance, but will not necessarily change the underlying dynamic of intrinsic motivation, skills and knowledge (e.g. capacity) necessary on a collective level to improve performance to the scale necessary.

The Relationship of Consequences (Response) and Supports

As discussed throughout this document, consequences cannot bring about the change envisioned by the policy makers without serious attention to support and capacity. The nature of consequences associated with the accountability system will discussed more completely in another section of this report. The discussion here focuses specifically on the relationship between consequences and supports.

If we are building a system that is truly focused on school improvement then it has to be more than consequence driven, in the typical sense of the word. Under performing schools should be provided targeted professional development to build the skill set of the teachers and administrators first, then if that is not enough then targeted intervention programs for students and technical assistance should be provided. When a student struggles we do not punish them, but engage them in a series of increasingly intense interventions designed to improve their performance. Why should it be any different for teachers and schools? But, just like with students any assistance provided to schools must be based upon data, and monitored for progress towards the target goals.

The Advisory Committee recommended that consequences for schools should be framed in the context of necessary and available supports and, as noted above, the performance designations are linked to increasing levels of required support. These consequences also should be linked with district accreditation, to the extent possible. While the general class of interventions, supports, and improvement goals should be outlined in the accountability system, the committee recommended that specific the accreditation targets and yearly or even multi-year improvement goals should be developed through a collaborative process between the school (district) and the state. This approach may help develop and support more internal capacity within schools and districts and lead to greater overall capacity statewide in the long term.

SECTION V: VALIDITY AND OTHER TECHNICAL ISSUES

This section of the report encompasses several key concerns related to the development and implementation of a standards-based accountability system. The major focus of this system is on the design of an accountability validity framework. Such a framework would guide the implementation of an evaluation of the accountability to make sure that it is functioning as intended and not leading to unintended negative consequences. Further, since the proposed accountability systems are based largely on standards-based assessment, this section begins with a discussion of the desirable characteristics and important validity concerns of both standards and assessments. Even though we present this section last, it is by no means least.

Standards: The Foundation of the System

This is called "standards-based" reform for a reason. The foundation of the system is the content standards that define what students are expected to know and do and the achievement (also called performance) standards that define how well students are expected to demonstrate understanding of the content standards. The goals of the accountability system implicitly invoke the use of content standards that will allow Wyoming policy makers to determine if, in fact, Wyoming students are reaching these goals. The first goal of having Wyoming become a national leader among states demands having a valid basis for making such comparisons. NAEP is typically used as a method of such judgments. There are many shortcomings with this approach, but the major problem is that NAEP results are not available at the district or school level. Further, even if NAEP was available at the district level, one would have to evaluate whether one district scored better than another because the higher scoring district's curriculum happened to match the NAEP framework more closely or because they were truly providing a better education. Using common standards would eliminate the first potential hypothesis to explain such score differences. Therefore, if policy makers and others want to compare Wyoming's performance with that of other states, having common standards makes such comparisons more plausible.

Wyoming policy makers indicated that having students leave Wyoming high schools college or career ready was a critical goal for implementing a comprehensive accountability system. Having content standards that define college and career readiness is essential if policy makers are serious about this goal and it is unfair to expect schools to meet this goal if there were no standards to serve as a guide for where educators need to aim. Further, the Select Committee members indicated significant concern with the levels of remediation required for Wyoming students in postsecondary institutions. Having content standards that help close the expectations gap between the end of high school and the beginning of postsecondary studies is critical so that students have a clear sense of expectations and educators have a similar understanding. Therefore, the Advisory Committee unanimously recommended that the State of Wyoming adopt the Common Core State Standards (CCSS). While there might be some legitimate concerns about the lack of control over the standards, the Advisory Committee felt that any concerns were far outweighed by both the high quality of the CCSS and for the reasons mentioned above.

Assessment Characteristics

The assessment system is the next leg of the standards-based accountability system and is critical in that it provides a great deal of the data for use in the accountability system. While a valid assessment system is necessary for having a valid accountability system, it is not sufficient because of the many other sources of data and decision rules that compromise accountability systems. Nevertheless, Wyoming policy and educational leaders should strive to have the highest quality assessment system possible to support accountability decisions. Therefore, we present specific considerations and criteria to bolster the likelihood that these assessments will support valid accountability decisions. We highlight only a few considerations here, because there are other important documents[13] that should guide the development of state accountability assessments.

Technical Characteristics

There are many technical characteristics of assessments important to the development of a valid accountability system all centered on supporting the validity of the inferences we draw from test scores, but we focus primarily on alignment, including rigor, reliability, and linking.

Alignment

Alignment, or the degree to which the test adequately measures the required content, is a critical technical issue for an accountability assessment. This falls under a fairness and transparency principle because those being held accountable (students, educators, school systems) should have a clear understanding of the knowledge and skills for which they are being held accountable. Alignment can get quite complex, but basically alignment is the degree to which test questions measure specific grade-level knowledge and skills represented by the content standards that teachers are expected to teach and students are expected to learn. Moreover, the degree to which the full set of grade-level standards is appropriately sampled by the assessment should be addressed in independent alignment studies. This “two-way” approach to alignment is important because many tests may claim alignment simply because the test questions match specific content standards even though important aspects of the standards are left untested.

If the content and performance standards are designed to represent college and career ready expectations, the assessments must also represent these expected outcomes. Any content and performance standards purportedly targeted to college/career readiness, and certainly the Common Core State Standards, demand demonstrations of complex thinking from students if they are, in fact, going to be declared “ready.” Therefore, the accountability assessments used in Wyoming must be able to measure students’ depth of understanding much more so than they do now. Doing this will require that a significant proportion of the test questions rely on formats

[13] The Standards are considered the “bible” and are more formally known at the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association). Additionally, the United States Department of Education’s Guidance for the peer review of state standards and assessment system is another important set of criteria, but based in large part on the *Standards*.

such as constructed response items and performance tasks where students are expected to generate their own responses and often provide a substantial explanation for their solution or response.

Reliability

Reliability, or the degree to which the test score can be expected to be consistent over time or over a different sample of items that represent the same domain, is a critical dimension of test quality, especially for accountability assessments. Reliability is simply the quantification of the error associated with any measurement. The *Standards* and the USED peer review guidance provide extensive detail about reliability and we will not go into great detail about reliability here. We briefly describe, instead, the importance of having a test that is fairly reliable across the full score distribution. If the main purpose of the assessment was to document whether or not students reached a specific cutscore (e.g., proficient), then it is really only important for the test to be reliable in the region of this cutscore. On the other hand, if the assessment is intended to provide useful information about all students and, most importantly, if it is designed to support growth measurement for students, the test should be fairly reliable throughout most of the score scale. This notion of reliability at specific scores is better discussed as the Conditional (conditioned on the particular test score) Standard Error of Measurement (CSEM). Tests used to support growth determinations do not have to possess equally low CSEM across the entire score distribution, but the CSEM at the high and low ends of the grade level achievement distribution should not be dramatically greater than the CSEM in the middle of the distribution. This requires that the test contain questions with a range of difficulty (the addition of open-response questions can help with this) and contains enough questions to support reliable inferences.

Scaling and Linking

While it is important that the test contains fairly low CSEM across the score scale, it is also important that the test does not have noticeable floor or ceiling effects. There is little doubt that fixed-form tests like most state assessments will have some students at the very highest and lowest scores possible. This does not pose a problem for accountability. However, it could cause challenges with growth determinations if there are noticeable percentages (e.g., 2% or more) of students scoring at the lowest or highest score on the tests. This will generally be more of a problem at the upper end of the performance distribution because as long as the test includes some multiple-choice questions, low achieving students are able to benefit slightly from chance and avoid scoring at the very lowest possible score. The range of item difficulty influences the highest and lowest possible scores, but decisions on how to scale tests can play a significant role as well. Scaling is the process of transforming the raw scores (the number of questions students answered correctly) to a scale that has more meaning across uses beyond that specific test form. Score scales are useful for communicating about acceptable levels of performance (e.g., proficiency) across test forms and occasions. Therefore, Wyoming should ensure that its tests are scaled appropriately to avoid floor and ceiling effects.

A meaningful and defensible score scale is certainly important to the success of the assessment and accountability system, but ensuring the specific test scores and/or achievement standards (proficiency) are comparable across test forms, especially across years is one of the most important aspects of the technical quality of accountability assessments. The process of linking, which represents a family of techniques that includes score equating, is how testing experts can state that a score of 200, for example, from 2010 has the same meaning as a score of 200 from 2011 even though the students from the two years did not take the same tests. The details of linking are too complex to discuss in this report, other than to say that linking and equating are complex enough that many testing contractors make errors in equating procedures that lead to unexpected declines or improvements in performance over years. The problems with linking are usually detected in the case of these unexpected score changes. What is more troublesome are the many cases where the score changes were not large enough to raise alarms. In this case, errors could accumulate over time and seriously threaten the validity of the accountability system. Therefore, as part of any testing contract, Wyoming must ensure that equating results produced by the main test contractor are verified by another equating expert either through a review of the procedures and results (a minimal level of quality assurance) to a full replication of the equating procedure (the maximum level of quality assurance).

Other Assessment Considerations

The testing industry has developed a sound knowledge base and set of procedures to ensure that the technical quality issues raised above are addressed appropriately. Of course, a third party, such as a high-quality technical advisory committee, must verify that these issues are, in fact, being addressed. There are other issues critical to the success of both the assessment and learning systems that are not often addressed in technical evaluations. The most important issue includes the role of a summative accountability assessment as part of a comprehensive assessment system. As we discuss in more detail below, assessments generally can serve one or two purposes well. If one tries to force an assessment to serve too many purposes, it means that it will not serve any of them well. So how then can an assessment provide both accountability and instructional information? It can't! Therefore, a comprehensive assessment system is required.

A comprehensive assessment system is one that includes assessments designed to serve multiple purposes (e.g., accountability/summative, formative/instructional, predictive, evaluative) with designs tailored appropriately for each purpose (Perie, Marion, & Gong, 2009). Again, this report is not the forum to go into great detail about comprehensive assessment system, but we discuss the role of a summative, accountability assessment in such a system. First, for a system to be comprehensive and function well, it must be coherent. What do we mean by coherence when it comes to the various assessments in a system? At a minimum the assessments must be targeted toward the same or at least purposely overlapping learning goals. Therefore, the summative, interim (if used in the system), and formative assessments must focus on the same learning goals or content standards. Of course, they can and should do so in different ways and with differing levels of granularity, but it must be clear that all assessments in the system are aiming at the same target. The summative, accountability assessment should go a step further and signal or represent the type and depth of learning we expect to see represented in curriculum, instruction, and in other assessments in the system. This signaling ensures coherence and helps make clear the expectation for learning, especially depth of learning, required in other parts of the system. To use a counter example, if the learning goals require students to solve complex problems and demonstrate a depth of understanding, but the accountability assessment only requires the demonstration of rote learning, it will not take long for the instruction to follow the accountability pressure and lead to teaching of low level outcomes only. Therefore, Wyoming's summative assessment system must include the types of problems and depth of understanding that we expect to see in high performing Wyoming classrooms.

The High School Assessment System

An important consideration for the development of a comprehensive and coherent assessment system is rapidly coming to light in discussions around the various assessment requirements for high school students. The recommendations from the Select committee as well as existing statute (SF 70) call for the use of "readiness" assessments including PLAN, EXPLORE, and ACT. Additionally, SF 70 calls for the use of end-of-course (EOC) exams in select high school course, to be determined in Phase II of the accountability system development. Further, the use of PAWS or replacement state assessment system will be required until WDE is able to receive permission from the U.S. Department of Education to replace the use of PAWS for NCLB accountability determinations with either components of the ACT Suite or select EOC exams. Finally, the proposed requirement for benchmark adaptive testing, depending on the grade levels ultimately determined, will add to the assessment burden at high school. Therefore, we recommend convening an assessment planning committee comprised of WDE assessment leaders, district and higher education representatives, and other stakeholders as appropriate to create a comprehensive and coherent high school assessment system.

Accountability Uses of Benchmark Adaptive Assessment

Wyoming's Senate File 70 authorized the use of benchmark computer adaptive testing to measure student longitudinal growth as part of the state accountability system. Apparently the intent of this provision was to broaden the accountability indicators beyond the state assessments and to use a measure of growth that essentially all school districts in Wyoming were already using. While this makes some intuitive sense, there are many concerns with this approach, specifically:

- Using an assessment for a purpose for which it was not designed,
- Concerns with the technically quality of the particular benchmark assessment, and
- The loss of any instructional value of the benchmark assessment by shifting to an accountability use.

These three concerns are all related and we briefly touch on each concern before offering some recommendations. Further, while the law (SF 70) did not name a specific assessment company, most involved in the legislation as well as observers acknowledge that Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) was the intended product implied by the legislation. However, our remarks below are directed toward interim/benchmark assessments in general.

Purposes and Uses

Perhaps the most important axiom in test design and evaluation is that the technical quality of tests can be evaluated only in the context of the specific purposes for which the assessment is intended to be used. For example, if an assessment is designed as an early warning indicator for how students are likely to perform on the end-of-year state test, then it must be validated for that purpose. Assuming the validity evidence is positive, that does not mean that the assessment is also valid for a different purpose such as program evaluation. Validity evidence would need to be gathered for additional purposes.

Most test vendors report that their products are useful for evaluating programs, informing instruction, and several other purposes. The validity evidence supporting any one of these purported purposes may be available, but since there has been little independent evaluation of almost all of these assessment systems and their reported benefits, one cannot demonstrate conclusively that there is evidence to support claims about such assessments for instructional, predictive, and/or evaluative purposes. On the other hand, accountability generally is not one of the stated purposes of benchmark/interim assessments, especially high stakes accountability. Therefore, little evidence would be available to support the accountability uses of any of these assessments. To be fair, some of these assessments possess some qualities that could potentially allow it to be used for accountability, but as described below, there are many shortcomings that could challenge the validity of such uses.

Technical Quality

What is a minimum level of reliability required for an assessment? This is a question that technical experts often are asked about assessments, but unfortunately, the answer is rarely clear cut. Essentially all experts will note that the level of reliability depends on the uses. If the assessment results are used to determine whether or not a student graduates from high school or whether a teacher is rated as effective or ineffective, for just two examples, then the test must be highly reliable. But if the results are just part of an ongoing set of information about how to inform/modify instruction, then the results of any particular assessment are not as critical and one could get by with lower levels of reliability. This is just an example, because the reliability of many interim assessments tends to be quite adequate.

Alignment, as discussed above, is critical for ensuring the validity and fairness of an assessment. We know of no independent alignment studies that have evaluated the degree to which any potential interim/benchmark assessments are aligned with Wyoming's content standards. For obvious reasons, independent alignment studies are much more credible than studies conducted by the test contractor. In fact, WDE and all other states were required to submit independent alignment evidence of the state assessment (PAWS) to the U.S. Department of Education as part of the federal peer review process.

The Center for Assessment has examined the alignment of state content standards (from other states) and provided technical advice on such studies in other states. In all cases, we found that claims that the test was fully aligned to the specific state's standards were considerably overblown. Further, all of the questions on most commercial interim/benchmark tests are multiple-choice. Many researchers and others have made clear that to appropriately represent the types of knowledge and skills called for by most state content standards (including WY), questions where students have to generate and supply their own responses (constructed and extended response questions) are needed. Therefore, the current crop of commercial interim/benchmark assessments will be unlikely to meet important "depth of knowledge" alignment requirements[14] as long as it remains a fully multiple choice based assessment. This problem will be exacerbated when Wyoming implements the Common Core State Standards

[14] We recognize that the SF 70 requirement to eliminate all constructed response questions from PAWS creates alignment problems from the state assessment as well.

(CCSS), because these standards require students to demonstrate considerably deeper understanding compared to most state assessments and this depth of knowledge should be assessed with items that require students to generate their own responses.

Perhaps the most significant concern with the technical quality of most commercial interim assessments is the generally low quality of the actual test items (questions). Adaptive tests are those where the computer program selects the items for the students to answer based on prior responses. The test stops according to a specific set of rules, but generally when the program has honed in on an accurate estimate of a student's achievement. Because it is critical to be confident in the pre-established item difficulty and the degree to which the items fit the theoretical model underlying the computer algorithm in this type of testing environment, the specific statistical properties of the item are often privileged over other aspects of item quality. In the past, commercial interim assessments have been criticized for low item quality (e.g., Shepard, 2006; Marion 2006), and while there is a chance that the item quality has improved, the constraints around item development for the huge pool of items required for an adaptive test, will likely mean that these assessments will suffer from lower quality items than custom designed large-scale assessments. Some might argue that these concerns about item quality are overblown, but if a test is to be used for accountability, especially educator accountability, policy leaders do not want to have to defend justifiable complaints about low quality test items.

Campbell's Law and Corruptibility

Much of what has been written above questioned the quality of the commercial benchmark assessments for many reasons, but mostly for their use as a potential accountability assessment. Even if we take at face value that these assessments provide instructional benefits—and there is no doubt that many school and district leaders report this to be the case—then a quick way to reduce any teaching and learning benefits of these assessments is to move them into an accountability context. This is not to say that assessments lose all instructional potential if they are used for accountability, but the fall-to-spring growth calculation used by some of these benchmark test vendors could easily be corrupted if educators are held accountable for these gains. Currently, educators have no vested interest in their students' performance on the fall test, but if educators and schools were accountable for the change in performance from fall to spring, they would actually have an interest in having their students perform poorly on the fall test so they could realize larger gains (all things being equal) on the spring test. This is just one example. There are many other possibilities for corruption and the loss of instructional usefulness if the benchmark assessments are employed for accountability purposes. To be fair, this caveat applies to any accountability design built on fall to spring measures of learning gains.

Recommendations

The following two major recommendations flow logically from the concerns expressed above.

- Do not use **any** commercial interim assessment as an accountability test.
- Allow districts to purchase interim/benchmark or formative products if they choose, but do not require the use of a single product. Districts should be able to choose based on needs and uses.

It should be clear by now that assessments designed for purposes other than accountability should not be used for accountability decisions unless the assessment can be validated for such uses. Interim and benchmark test vendors are generally not very specific about the intended purposes of their assessments in order to appeal to as broad a market as possible, but even still, very few, if any, interim/benchmark tests are marketed as accountability tests and validated for these uses. Further, by using such tests for accountability, the users run the considerable risk of giving up on the purported instructional benefits of these assessments. Therefore, there is little rationale for having the State support (i.e., pay for) the using of a common benchmark or interim assessment product.

The second recommendation follows directly from the first. The policy makers should certainly allow district leaders to use their block grant funds to purchase an interim/benchmark assessment program, support formative assessment initiatives, or create their own common assessment program. There is a fair body of research supporting the use of formative assessment practices for improving student learning, but there no such corpus of research supporting the use of interim/benchmark assessments for these purposes. Considering this lack of research, it makes little sense to advocate a specific interim assessment product or a particular model of use (e.g., administered three times per year). Rather, districts should be free to select the model that they think will work best for their context and needs, evaluate the efficacy of such a model in their districts, and adjust the testing program if necessary.

Evaluation of the Accountability System

In addition to evaluating the technical characteristics of the assessments, it is critically important to evaluate the accountability system. We cannot overstate the importance of a comprehensive investigation prior to implementation and ongoing monitoring and support following implementation in order to maximize the likelihood that the state's objectives will be met.

Following, we present key claims that should be investigated in the evaluation process along with exemplar studies to inform each. Although not comprehensive, these components are intended to capture the core areas that should be examined to evaluate the suitability of the model.

Evidence Supports Claims in the TOA

This claim addresses the supports and structures that must be in place to bolster the integrity of the information in the model and to improve the likelihood that actions based on information derived from the accountability model will promote intended outcomes.

This broad claim connects to many aspects of Wyoming's education system including:

1. The content standards and resulting curricular frameworks are designed around a credible learning progression and they represent the knowledge, skills, and abilities necessary to promote college or career readiness.
2. State assessments provide reliable and valid scores.
3. Academic growth information based on state and/or other assessments is credible and technically defensible.

4. Educators have access to the right information and have the knowledge, skills, and support necessary to improve student learning.

Results are Reliable

Reliability refers to the consistency or stability of a measure. In this case, we are interested in the reliability of the measures of schools or teacher/leader outcomes. Reliability is challenging in this context due to the error in both achievement measures and growth measures.

Additionally, reliability is impacted by sampling error. Sampling error refers to fluctuations in school or class outcomes scores that can be unrelated to actual school performance. In fact, Hill and DePascale (2002) emphasize that sampling error, "contributes far more to the volatility of school scores than measurement error." Sampling error can work to both the advantage and disadvantage of schools on reported accountability determinations, but the goal is still to minimize the effects of sampling error on school results.

There are multiple statistical approaches to evaluating the reliability of school or class determinations. However, at a minimum it is advisable to track the consistency of outcomes for various levels (e.g. schools, subgroups) within and across years. Although not without exception, it is expected that results will be well correlated for similar school types within year and for the same schools across years. Dramatic shifts in either classification of schools or characteristics of the distribution will signal a troubling lack of stability that will erode the credibility of the outcomes.

Results are Valid

If reliability addresses the extent to which the model provides a consistent answer, validity asks, "Is the answer correct?" Stated another way, to what extent are the results credible and useful for the intended purposes? At a minimum, an investigation of the validity of the model should address the following:

1. Is the model appropriately sensitive to differences in student demographics and school factors?
2. Are the results associated with variables not related to effectiveness or generally those not under the control of the school, such as the socioeconomic status of the neighborhood?
3. Are the classifications credible?
4. Are negative consequences mitigated?

The first question addresses the extent to which the model differentiates outcomes among schools and/or classes. A model in which very few schools differ with respect to results (i.e. all ratings are high) will likely be out of sync with expectations and the credibility of the results will be suspect. Therefore, it is important to examine the distribution of results to determine if the outcomes are sensitive to differences and if the dispersion is regarded as reasonable and related to expected differences in school quality as documented from other means.

Second, it is important to examine the distribution of scores with respect to variables that should not be strongly associated with outcomes. For example, if there is a strong negative relationship between student poverty and school scores (i.e. lower poverty= higher scores) this suggests that effective schools are only those in which relatively affluent students are enrolled. Similarly, if there is a strong positive relationship between a student's prior year achievement and a rating of educator effectiveness, this indicates that the most effective teachers are those in classrooms where the students started out as high performing. Such findings are implausible and erode credibility of the model.

The third question calls for examination of classifications with respect to external sources of evidence that should be correspondent with quality. For example, one would expect a higher percentage of teachers who have been certified by the National Board of Professional Teaching Standards to be classified as effective compared to those who are not. Similarly, high schools with higher graduation rates or higher college-going rates should, in general, receive more favorable outcomes that schools struggling in this area. It should be clear that if the school accountability model is intended to identify and reward those schools that are preparing students for college and career, the validity evaluation will be incomplete without including data that reaches beyond K-12 and provides an indication of the post-secondary outcomes for graduates.

Finally, a validity evaluation should address the extent to which unintended negative consequences are mitigated. If potentially troubling consequences such as narrowing the curriculum, reduced professional cooperation, educator transition/attrition, or cheating on standardized tests occurs, the validity of the system is threatened. Some of these threats could be examined via survey data or focus groups, while others may be explored with extant data. Importantly, ongoing initiatives to gauge the extent to which positive outcomes outweigh potential negative side effects will bolster the consequential validity of this initiative and provide a mechanism to promote continuous improvement.

References

Conley, D. (2005). *College knowledge: What it takes students to succeed and what we can do to get them ready*. San Francisco: Jossey-Bass.

Data Quality Campaign. (2010a). Strengthening the teacher-student data link to inform teacher quality efforts. Retrieved from: www.DataQualityCampaign.org/resources/947.

Data Quality Campaign. (2010b). Developing a definition of teacher of record. Retrieved from: http://dataqualitycampaign.org/files/Teacher%20of%20Record.pdf.

Domaleski, C.S. (2011). *State End of Course Tests: A Policy Brief.* Paper commissioned by the Council of Chief State School Officers Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards.

Domaleski, C.S. & Hill, R. (2010). Considerations for Using Assessment Data to Inform Determinations of Teacher Effectiveness. Retrieved from: http://www.nciea.org/papers-UsingAssessmentData4-29-10.pdf

Betebenner, D.W.(2009). Norm- and criterion-referenced student growth. *Educational Assessment: Issues and Practices, 48* (4),pp. 42-51.

Betebenner, D.W. & Linn, R. L. (2010). Growth in student achievement: issues of measurement, longitudinal data analysis, and accountability. Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda: Center for K-12 Assessment and Performance Management. Retrieved from: www.k12center.org/rsc/pdf/**BetebennerandLinn**PolicyBrief.pdf

Briggs, D. & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness ranking of Los Angeles Unified School District teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved from: http://nepc.colorado.edu/publication/due-diligence

Fink, S. & Markholt, A. (2011). *Leading for instructional improvement: How successful leaders develop teaching and learning expertise.* San Francisco: Jossey Bass.

Glazerman, S., Goldhaber, D., Loeb, S., Raundenbush, S., Staiger, D. and Whitehurst, G. (2011). Passing muster: Evaluating teacher evaluation systems. Brown Center on Education Policy at Brookings. Retrieved from: http://www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx.

Hill, R.K., & DePascale, C.A. (2002). Determining the reliability of school scores. Portsmouth, NH: The National Center for the Improvement of Educational Assessment Inc. Retrieved from: www.nciea.org

Linn, R. L. (2008). Educational accountability systems. In The Future of Test Based Educational Accountability, pages 3–24. Taylor & Francis, New York.

Marion, S.F. & Buckley K. (2011). Approaches and considerations for incorporating student performance results from "Non-Tested" grades and subjects into educator effectiveness determinations. Retrieved from: www.nciea.org

McCaffrey, Daniel F., Daniel Koretz, J. R. Lockwood and Laura S. Hamilton (2003). Evaluating Value-Added Models for Teacher Accountability. Santa Monica, CA: RAND Corporation. Retrieved from: http://www.rand.org/pubs/monographs/MG158

National Research Council. (2010). Getting value out of value-added. H. Braun, N. Chudowsky, and J. Koenig (eds.). Washington, DC: National Academy Press.

Perie, M., Marion, S.F., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28, 3*, 5-13.

Raudenbush, S. (2004). Schooling, statistics, and poverty: Can we measure school improvement? (Technical report). Princeton, NJ: Educational Testing Service. Retrieved from: www.ets.org/Media/**Education**_Topics/pdf/angoff9.pdf

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 29(1):103–116. Retrieved from: www.ucce.ucdavis.edu/files/workgroups/6798/**Rubin**EtAl.pdf