APPENDIX B: ACCOUNTABILITY CHALLENGES

Accountability Challenges

Overview

Wyoming's legislatively created and mandated accountability system has been a major issue over the current superintendent's term. Wyoming's accountability system has been heavily influenced by the requirements set forth by the US Department of Education (USDE). The legislature tailored the system first to meet the criteria for the Race to the Top (RTTT) competitive grant program and then to comply the No Child Left Behind (NCLB) waiver requirements.

PAWS

Wyoming developed an assessment system for NCLB that is well described in chapter 7 of *Unlearned Lessons* (Popham, 2009). That system was based on a state developed test Proficiency Assessment of Wyoming Students (PAWS). The development was overseen by the test vendor, the Wyoming Department of Education, and the Wyoming Technical Assistance Committee (TAC). The TAC is a committee made up of experts in assessment paid by the USDE to provide expertise not held by the WDE. Wyoming's TAC sought to influence the design of PAWS so that it would be instructionally supportive, or in other words would be helpful in improving instruction of Wyoming students in the areas measured by PAWS. The history of PAWS shows that the test was often changed for various reasons prior to 2011.

- 2006- First administration of PAWS in spring.
- 2007- PAWS administered twice with banking of scores.
- 2008- PAWS administered once and Science was added to the areas measured by PAWS.
- 2009- The scoring of writing on PAWS changed from a six trait rubric to a four trait rubric.
- 2010- The administration of PAWS failed and the results could not be used for accountability required by NCLB. A waiver for the failure of PAWS was obtained from USDE and claims were made against the vendor for the failure resulting in a settlement in April, 2011 (Moore, November 19, 2010).
- 2011- PAWS moved from a hybrid, computer and paper test to just a paper test.

While the changes in PAWS frustrated teachers, then the Wyoming Accountability in Education Act work began. The impact to the instruction of students in Wyoming by changing assessments cannot be measured, but certainly the impact to teachers has been significant. Below I discuss some of the inherent problems in the current assessment and accountability system.

Wyoming Assessment

Wyoming has never had a complete set of peer reviewed assessments that meets the requirements of NCLB. The system was very close to approval in spring 2011. It is left to the reader to determine if the current system is best for Wyoming students or if further changes in assessment are in the best interests of the state and the public education system. Such

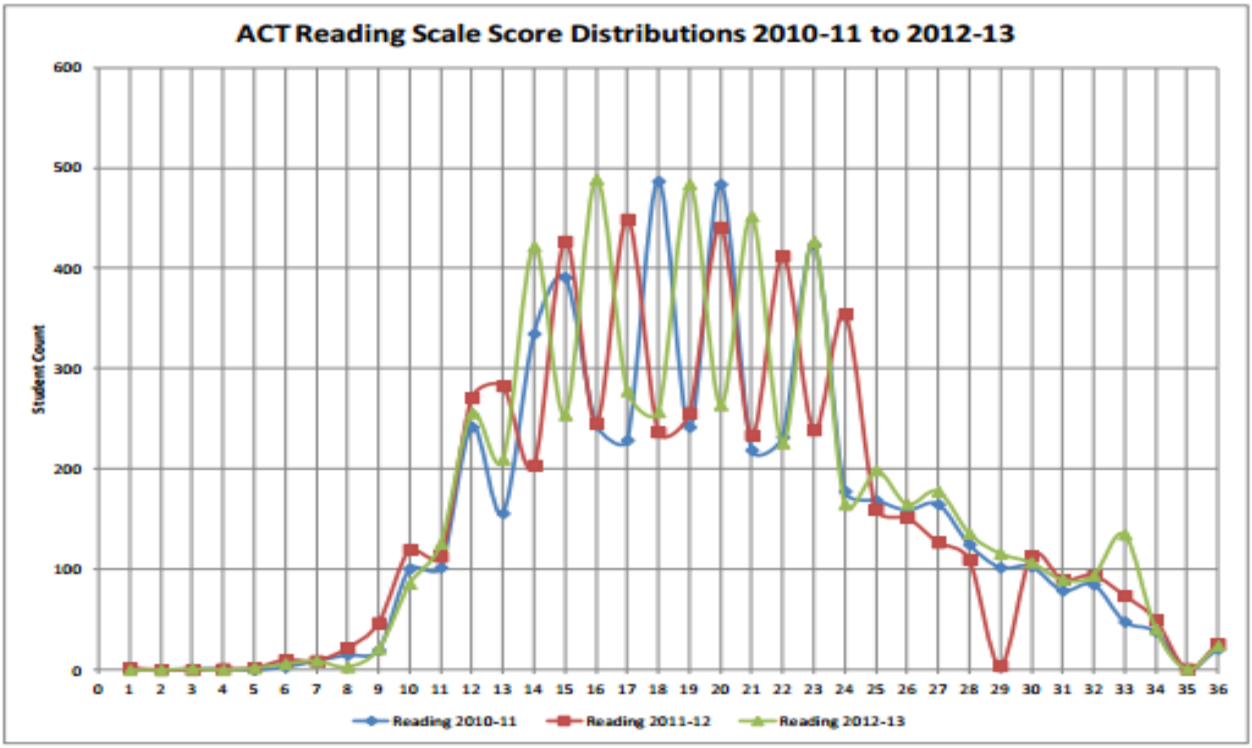
decisions should be made only after thoughtful deliberation as changing assessments will continue to undermine teachers and instruction. Not changing assessments in the current system leave problems in place that will likely be the subject of federal mandates or state legislation. Consider the following issues:

- ACT™ with writing is not aligned with Wyoming Content and Performance Standards. It is used to predict future success in college but not to measure how successfully the Wyoming standards have been taught.
- ACT™ Scoring for Wyoming accountability system differs from traditional ACT™ scoring which is confusing to educators (in-depth discussion below).
- ACT EXPLORE™ and ACT PLAN™ are used as accountability assessments for 9th and 10th grade students. These assessments are obsolete and being phased out of the ACT Suite for business reasons.
- ACT has not been approved as an assessment tool by the USDE for Wyoming AYP calculations.
- SAWS is not a writing test, but rather a response to reading test.
- SAWS is part of Wyoming accountability, but is not part of federal accountability.
- PAWS-ALT has been used for 11th grade and has had large changes since 2012.
- WY-ALT will replace PAWS-ALT, but the science standards will not match Wyoming standards.
- WY-ALT will yield English Language Arts (ELA) scores while PAWS will measure only reading for federal accountability.
- NWEA MAP is legislated for all districts to use as a district assessment, yet the white paper that was part of the WAEA in 2012 discourages such benchmark assessments.
- PAWS is entirely multiple choice, but the removal of constructed response items decreases the reliability of the assessment.
- The majority of the assessment system is multiple choice, but has the stated goal to prepare students to be college and career ready… all of our research efforts have failed to discover a multiple choice career.

The mixed messages from the accountability legislation and implementation of that legislation have left educators wondering what changes will come next.

Specific Measures of Accountability and Concerns:

ACT™ with writing: The ACT™ will be scored differently this year for accountability purposes. The reason has to do with the test itself. There are some basic psychometric rules in play with the design of all assessments and they are at play in the scoring of this well known and frequently used assessment for college readiness. The first rule is that subsections of a test are always less reliable than the complete test. The best example of this is the reading test for ACT™. Below is a graph that shows the distribution of the numbers of students for each scale score for reading with each colored line representing a year of administration in Wyoming.



The chart shows that there are significant changes year to year in the counts of students for scores between 13 and 24. In other words, if a proficiency cut score is placed between 13 and 24, it is likely that the numbers or percentage proficient will change radically year to year with minor changes in actual instruction or student knowledge.

This problem was presented to the Wyoming TAC as well as ACT™ in hopes that a solution could be found to provide a more consistent means of evaluating student proficiency for the Wyoming accountability model. The result is the Wyoming Scale Score for ACT™. The new scale score will present some confusion to stakeholders. Consider the following observations. Often the Wyoming Scale Score of a student may be lower than students who achieved a higher score on the normal ACT™ scale or vice versa. In the extreme cases, over a thousand students may score better on the Wyoming Scale Score and worse on the ACT™ scale score.

READING WY_SCALE_SCORE	READING ACT™_SCALE_SCORE	# students with lower WY Scale Score and higher ACT™ Scale Score	# students with higher WY Scale Score and lower ACT™ Scale Score
123.472	12	1022	0
122.144	20	0	1642

The student who has an unrounded score for reading on the Wyoming Scale Score of 123.472 has 1022 students who scored lower on the Wyoming Scale Score but all of those students scored 13 or higher on the ACT™ scale score. This represents about 18% of 11th grade students who took the test. Conversely nearly 29% of the 11th graders had a reading score on the Wyoming Scale Score greater than 122.144 with a lower ACT™ scale score than 20. Additionally there were 178 students who scored the minimum score of 21.42 (unrounded) and had scores on the ACT™ scale score ranging from 3 to 17. In fact, any Wyoming Scale Score will correspond to several ACT™ scale scores.

This phenomenon will confuse any educator in Wyoming. Overall scores as measured by ACT™ have increased and proficiency in Wyoming is lowered leaving stakeholders scratching their heads.

The explanation for this dichotomy is that the Wyoming Scale Score is **not** comparable to the ACT™ scale score. The same test was scored for each student in two different ways. ACT™ is normally scored using what is known as one parameter Item Response Theory (IRT). The Wyoming Scale Score uses a three parameter IRT method. The difference is that a one parameter measurement relies only on correct scores. A three parameter relies on correct and incorrect scores as well as the probability of student guessing. The three parameter measurement produces a much smoother curve than the one parameter and hence, more consistent proficiency scores.

Of course, the actual scoring procedure and exact methods to accomplish the scoring is not known by anyone but the vendor, ACT™. Because the vendor operated in the competitive testing market with this particular test, the scoring will be considered proprietary and will be likely kept secret. The best that educators in Wyoming can do to prepare students for the ACT assessment is to simply teach their students well.

Graduation and 9th Grade Credit Accumulation: Graduation rate is a prominent indicator for the accountability model. Should it be? In the last October 15, Report to the Legislature (2013), a section was devoted to enrollment. Graduation rate in Wyoming is a slow calculation. The measure starts in the fall of 9th grade where enrollment jumps on average 4% above that in eighth grade. Students come and leave the school over the next four years until a Spring or Summer graduation. The WDE then takes up the mechanics of calculating graduation rate to be completed in the spring of the next year. The PJP meets in September and the model is calculated in October, slightly more than 5 years from the initial measurement. The issue of timeliness matters in an accountability system. Additionally, we have 9th grade credit accumulation as part of the accountability model. According to a study done in Natrona County School District #1:

> "This study also found that losing one or more credits attempted during grade 9 was a powerful predictor of dropping out of school. Therefore, reducing the percentage of students who lose one or more credits during grade nine should also

become a district goal. Thoughtful problem solving at the district and school level to develop a plan to improve success in grade nine classes should occur."

The use of both 9th grade credit accumulation and graduation rate might be measuring the same thing in differing ways but with differing timeframes. The 9th grade credit accumulation is a predictor of dropping out of school and graduation rate is a confirmation of that predictor four years later. The question is whether these measures are appropriate for accountability. What is the behavior or practices that would better increase the completion of high school and are these the measures that accomplish this?

Measurement of the Hathaway Curriculum:

Measuring a curriculum on a state basis may be the same as installing a statewide curriculum. The state sets the standard. The state measures the standard. Local control is supposed to set the curriculum to meet the standards. When the state measures the Hathaway curriculum, then the state may be infringing on the curricular decisions of districts, schools and students. This is not a criticism of the Hathaway scholarship curriculum that is currently a choice for students, but of the measure that will likely force students to take a particular path in high school that may not be in their educational interest.

Summary of Accountability:

It is important to note that there are two important assumptions within the Wyoming accountability model. The first assumption is that the average errors in the measures for accountability are zero. The second assumption is that the statistical distributions are all normal distributions.

Assumption—Error is zero:

Error is an important element in the model. If error causes a school to be evaluated incorrectly, then the consequences for the school are large. Consequently, the state can experience financial consequences if error causes a school to consume resources that should have been applied to another school. The basic equation is below and applies to all the measures in the Wyoming accountability model.

$$\text{True Score} = \text{Measured Score} + \text{Error}$$

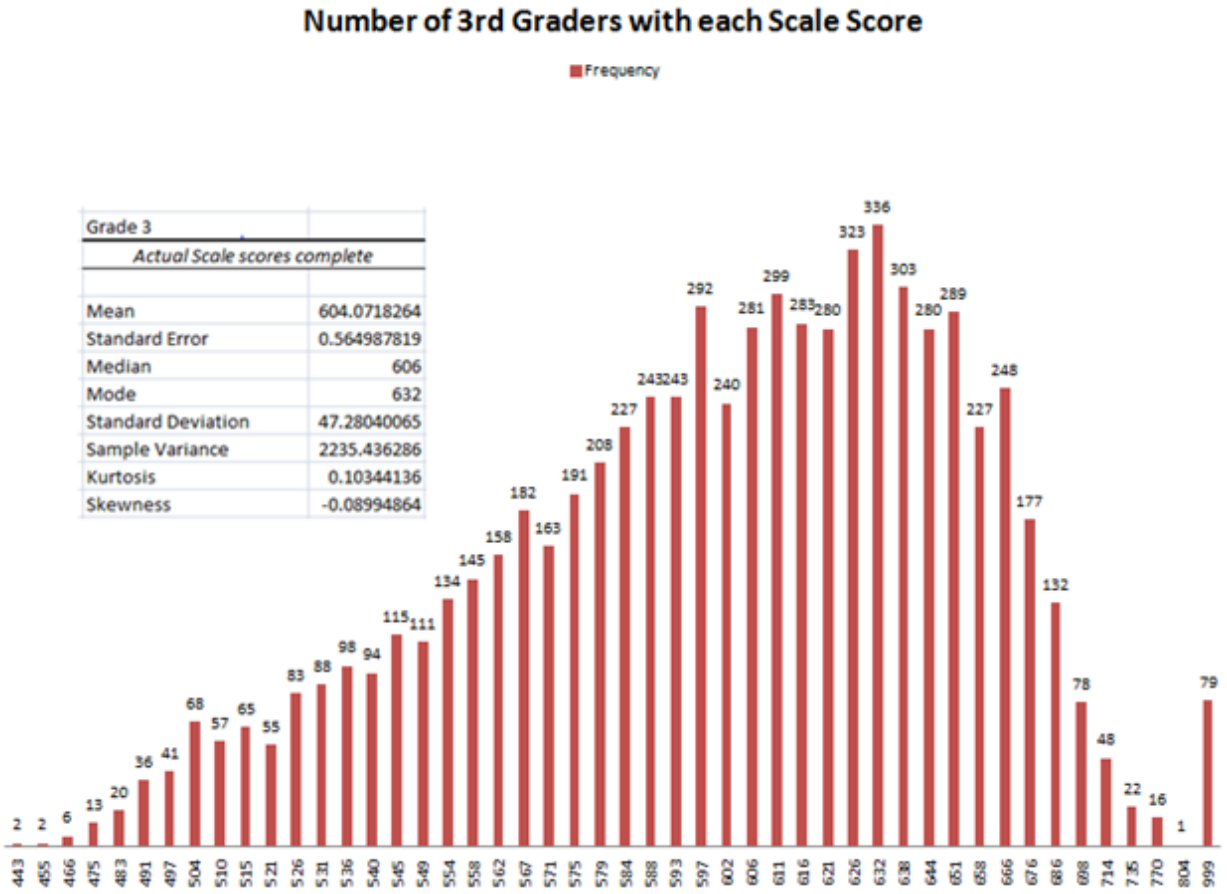
This fundamental equation is to be applied to all measurements and often people will assume that on average the error function will be zero. However, if the error is large or if the error function is heavily skewed, then this assumption is flawed.

Assumption—Normal Distribution Statistical Analysis Applies:

In 1973, Fischer Black and Myron Scholes published a model which estimates the price of financial options over time. Robert Merton then expanded upon that model and coined the name Black-Scholes options pricing model. In 1994, Robert Merton and Myron Scholes started Long-Term Capital Management (LTCM) a hedge fund. In 1997, Merton and Scholes were awarded the Nobel Prize in economics. The hedge fund had produced returns, after fees, of

21%, 43% and 41% by 1997 using methods that were based on normal distribution statistics. They were the smartest men in the room. In 1998, the hedge fund lost $4.6 Billion, in part due to Skewness and Kurtosis Risk. They were trading on distributions that were not normal using norm referenced statistics.

There are no normal distributions in the assessments of the WAEA. The distribution for ACT™ is clearly not normal. It is commonly referred to as the "Hair on Fire" distribution because of the random spikes in the number of students achieving a particular score year to year. PAWS distributions are skewed and have fat tails to the left. The mode is always greater than the median and the median is always greater than the mean. An example of a PAWS distribution is below.



But you can see that the two main assessments are clearly not normal distributions. The model for WAEA relies on regressions and statistical analysis that assume normal distributions. They assume that the error functions are normally distributed and centered at zero. Given the fundamental distributions in our assessments, it is likely that those assumptions are wrong. Are the foundational assessments that contribute to the high-stakes decisions of the accountability model subject to the same risks that required the Federal Reserve to bail out LTCM?

The model may be flawed. But the question is really if it produces information that results in wrong decisions. We cannot know. Not at this time. Dr. Joe Ryan, who previously served on

the Wyoming TAC, said that "if you are going to use an assessment for a purpose that it was not intended, then you should test it for three years". It was wise advice for assessment. But what would be the advice for accountability? Perhaps Wyoming should attempt to validate the model. Perhaps Wyoming should compare the model to the logical determinations that would have been used a number of years ago. It is likely that schools that were in trouble in 2011 are the same schools in trouble using the WAEA model. How much money has Wyoming spent on a model to make decisions on how to deploy a system of support, which makes essentially the same decisions that would have been made prior to WAEA? Certainly some expert oversight would look closer into questions concerning the high stakes model produced for WAEA.

Consider the conclusions and supporting evidence from the State Board of Education in Vermont. In the statement, there were eight well presented guiding principles. I encourage everyone to read the Vermont State Board of Education Statement and Resolution on Assessment and Accountability that is attached in the Appendix. One point made there is that tests are designed for a purpose. These tests lose validity and fairness when they are applied to purposes for which they are not intended. No test has been designed for the purpose of evaluating the performance of a school in all areas of instruction. No test of students has been designed for the purpose of evaluating teachers.

What is the purpose of the Wyoming accountability system? The first stated goal is to be a national leader in education. The unstated goal is to meet requirements of the US Department of Education. These two goals are in serious conflict. Consider the performance of the District of Columbia schools, the only school district under control of congress.

- 2013, fourth grade reading NAEP scale score of 206, lowest score amongst all states
- 2013, eighth grade reading NAEP scale score of 248, lowest score amongst all states
- 2013, fourth grade mathematics NAEP scale score of 229, lowest score amongst all states
- 2013, eighth grade mathematics NAEP scale score of 265, lowest score amongst all states
- 2012, four year cohort graduation rate 59%, lowest score amongst all states

Perhaps, Washington DC should not be the place we look to for education policy or accountability.

Considering the above data and circling back to the conclusions reached in Vermont, should Wyoming change the focus of the accountability system from tests not being used for the purpose that they are designed, to focus on student learning .

We have what we currently have for the accountability model formed by the legislature. Stakeholders need to know that the model is subject to error. They need to examine the decisions made by the model and question if the decision could have been wrong based upon the standard errors that come with the measures in the model.

The Legislature and the contractors involved with the creation of the model should explore the measurements to know if they are accomplishing what was intended. If a measure is not timely, perhaps there is a better measure. If a measure has a high amount of error, then it might be improved. Overall, the accountability model should be subjected to testing using the known errors in a simulation to prevent capricious decisions and know the impacts of error. Most of all the model should be one that supports learning and minimizes impacts to instruction.