EVALUATION AND CONTINUOUS IMPROVEMENT PLAN FOR THE WYOMING ACCOUNTABILITY EDUCATION ACT (WAEA) SCHOOL ACCOUNTABILITY SYSTEM

A Progress Report from the Wyoming Accountability Advisory Committee to the Joint Education Interim Committee: Part 2

Judy Catchpole, Chair

Report[1] prepared by:
Joseph Martineau and Scott Marion
Center for Assessment

October 8, 2018

[1] This report draws heavily on D'Brot, Lyons, and Landl (2017).

Advisory Committee Members

Judy Catchpole, Chair of the Advisory Committee, Business representative

Tony Anson, Principal, Riverside School, Big Horn County School District #4

Sue Belish, State Board of Education

Diana Clapp, Superintendent, Fremont County School District #6

Kris Cundall, Principal, Walnut Elementary School, Sweetwater County School District #1

Tess Hopkin-Egger, Reading Success Teacher, Livingston Elementary School, Park County School District #6

Mary Kay Hill, Governor Mead's Policy Director and Education Advisor

Sean McInerney, Accountability Supervisor, Wyoming Department of Education

Jody Rakness, Assessment Director, Washakie County School District #1

Terry Snyder Superintendent, Fremont County School District #25

Janine Bay-Teske, School Board Member, Teton County School District #1

Brandon Cone, Secondary Engineering Design Teacher, Campbell County High School, Campbell County School District #1

Executive Summary and Recommendations

This document describes an evaluation and continuous improvement plan proposed by the Wyoming Accountability Advisory Committee to support the Wyoming school accountability systems. The Advisory Committee prioritized certain studies in the plan based upon the need, potential impact, and likely burden on school and district staff members. We provided rough estimates of these costs in the text and provided details in the various appendices and we noted that these costs can be spread over a few years. The rough estimate for the multiple studies within the highest priority category of studies is approximately $250,000. We provided details on the study designs so that we could estimate the scope of work necessary to conduct each study and so that we could estimate a reasonable budget for conducting the particular studies. These study designs should not be viewed as fixed. Rather, they should serve as estimates of the effort and cost necessary to conduct these or similar studies. The Advisory Committee strongly recommends allowing Wyoming Department of Education (WDE) to make incremental progress on the research agenda to build a growing base of validity evidence in support of the accountability system and to provide information for how best to continue to improve the accountability system.

Introduction and Purpose

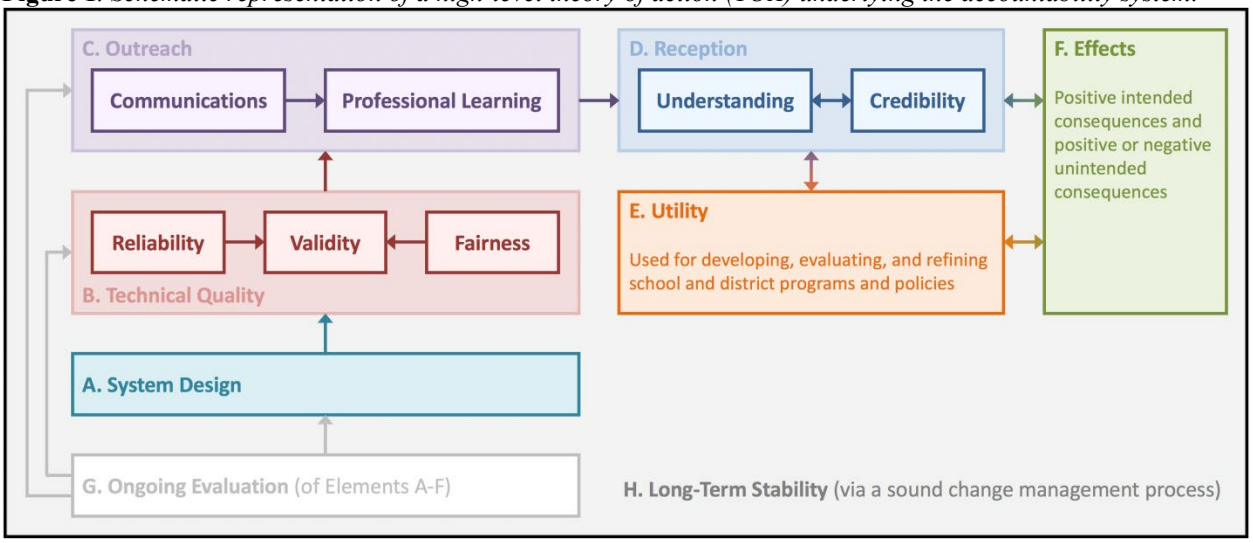
The Wyoming Accountability Advisory Committee (Advisory Committee) was charged with designing an accountability system for the Wyoming public school system. Further, the 2017 interim report of the Advisory Committee recommended that the Wyoming Department of Education and the Advisory Committee develop a plan for evaluating the accountability system and creating a continuous improvement framework. The overall goal of this exercise is to evaluate whether and to what degree the accountability system is working as intended. In this case, "working as intended" means two things. First, is the system producing results that are credible and defensible by, for example, identifying the "right schools" for support and assistance? There is no point in having a statewide accountability system that is unable to perform that critical first task. However, states do not create accountability systems simply to identify the "right schools" for particular support and services. The ultimate goal of creating policies and practices for a school accountability system is to improve the outcomes for students and schools. Therefore, this is the second and ultimately most important purpose of school accountability systems. The evaluation plan described in this document was designed to address both of these purposes. Additionally, the evaluation approaches described in this document are not meant as formal summative determinations regarding the quality of the accountability system. Rather, the Advisory Committee views the suggested studies and analyses as ways to support a continuous improvement process during the life of the accountability system.

We begin the report with a description of a high-level theory of action for Wyoming's school accountability system. Next, we outline the various system components the Advisory Committee recommends including in an evaluation plan. Third, we build out these components by outlining potential evaluation questions and study designs. Finally, we include a discussion of the costs necessary to conduct each study. We know that such costs are rough estimates at best, but we argue that they provide a good starting point for Wyoming policy makers.

Theory of Action

The Advisory Committee described a high-level theory of action for how the accountability system is intended to result in improved outcomes for Wyoming students. That high-level theory of action is displayed schematically in Figure 1 on the next page.

Figure 1. *Schematic representation of a high-level theory of action (TOA) underlying the accountability system.*



Note: *A → B does not mean A causes B, but that an effective approach to A improves the likelihood that B will be effective.*

The school accountability system has multiple levels of components. The overall score is composed of *indicator categories*, which are composed of one or more *indicators*, which are in turn composed of one or more *measures*. The Advisory Committee developed the measures to be actionable for educational improvement efforts. Therefore, the measures and indicators provide the data most likely to have utility for developing, evaluating, and refining school and district programs and policies, so the majority of the evaluation plan focuses on these components of the system.

The Advisory Committee discussed and deliberated the major elements of a Wyoming accountability system that should be addressed as part of a comprehensive evaluation plan and agreed that the following elements should be present in an evaluation plan:

- ✓ System design
- ✓ Technical quality
- ✓ Outreach/communication
- ✓ Reception
- ✓ Utility
- ✓ Effects
- ✓ Ongoing evaluation
- ✓ Stability

We expand on each of these elements below.

<u>System Design</u>

There is a long history of thoughtful design with the Wyoming Accountability in Education Act (WAEA) and that experience was used to support the design of the ESSA accountability system. An effective design based on thoughtful deliberations with a range of stakeholders is a necessary requirement for having a high-quality accountability system. Further, the design documents

must be sufficiently specific to guide business rule development so that analysts do not have to anticipate value-based decisions.

Technical Quality

The Advisory Committee focused on three main aspects of technical quality: validity, reliability, and fairness. **Validity** means that measures and indicators used in the system lead to inferences that are supported by logic and evidence. This means that input data (e.g., test scores) are strong proxies of the things we care about measuring (e.g., mathematics achievement and growth), the business rules closely follow the intended system design description, and the things being measured are under the control, at least partially, of the school. **Reliability** is based on sampling theory and is essentially a measure of the consistency and uncertainty with certain measures or indicators. In the case of the school accountability system, measures used in and created by the system are consistent and accurately calculated. A first condition of consistency is that the business rules allow for near-perfect replication of the accountability system results. Once that condition is met, reliability analyses examine such things as the year-to-year consistency of school results across years. **Fairness** is important from both a social justice and a system credibility perspective. For example, if all of the schools that perform well in the accountability system are located in wealthy neighborhoods and all of the schools that perform poorly have a high percentage of economically disadvantaged, then an accountability system is not really needed. All that is needed is a census map in this case. A fair accountability system means that measures used in and created by the system do not inappropriately and disproportionately identify schools serving different populations.

Outreach/Communication

The Advisory Committee spent considerable time discussing the importance of ensuring that all of the various stakeholders understand the rationale behind the accountability systems, the designs, and, most importantly, what the results mean. This requires the state to develop and implements a communication plan that is clear and effective in motivating stakeholders to review accountability reports and learn how to understand and use them.

Another critical aspect of the communication and outreach efforts involves producing professional learning materials that are clear and effective in helping local policymakers understand and use the information in accountability reports. This includes designing reports that support allow for users to learn accurately about school performance based on a clear and intuitive structure for navigating and laying out accountability reports. This also includes providing easy-to-access interpretative tools to allow users to further their understanding of the report and associated information. Outreach and communication are cornerstones of ensuring public credibility of the accountability system.

Reception

Reception is the next logical step to the outreach/communication work described above. Understanding the information contained on the accountability reports is the required first step to using the information to inform the development, evaluation, and/or refinement of school and/or district policies. In order for stakeholders to use the reports for making educational decisions, they must first see the reports as accurate reflections of school and district characteristics and as useful for developing, evaluating, and refining school and district programs and policies. As

credibility is built, more stakeholders are motivated to gain understanding, creating a cycle of increasing understanding and credibility.

Utility

Henry Braun, noted measurement and accountability expert, has said that utility is the most important criterion for accountability systems. Utility refers to the ways in which the design and results support stakeholder actions to bring about the desired actions and hopefully improvement. When local stakeholders understand and see accountability reports as credible, it can help them use accountability reports to meet their needs for developing, evaluating, and refining school and district programs and policies. As more stakeholders use the data to inform their decision-making, the system gains more credibility, motivating more educators to gain understanding of accountability reports, creating a cycle of increasing understanding, credibility, and data use.

Effects

When local stakeholders use accountability data, it can help to achieve the intended effects (consequences) such as improved student outcomes via improved teaching and learning practices. Systems with a high degree of utility should result in positive unintended consequences (e.g., creating a culture of data-based decision making) with minimal unintended negative consequences (e.g., inattention to non-tested grades/subjects).

Stability

The Advisory Committee has been emphasizing since its inception the importance of system stability. The Wyoming Assessment Task Force of 2015 echoed this sentiment in terms of Wyoming's statewide assessment system. The challenges of translating policy into classroom practice are well known so it is very challenging to realize system goals when policies change every few years. Therefore, an accountability system must remain as stable as possible in the long term, with minimal changes only as necessary. All potential revisions are made based on an established process to ensure that state accountability policy continues to support evaluation and continuous improvement of local policies and programs.

Recommended Evaluation Questions and Associated Studies

In this section we develop one or more individual research questions for each aspect of the theory of action. For each research question, we present a narrative plan for answering the question using at least one approach described in the preceding section, followed by a description of the level of effort involved in completing the study and an estimated budget required for the study[2]. We label the recommended studies alphabetically (A through K), but note that several research questions do not have recommended studies at this time. A brief outline is provided below.

[2] Estimated costs are divided into two categories: time and materials/expenses. Daily rates are derived by obtaining a median annual salary for an employee in an appropriate job classification from indeed.com. That salary is then doubled (to represent benefits and corporate profit), divided by 260 working days per year (a typical work year with holidays plus two weeks' vacation), and rounded up to the nearest $50/day to avoid underestimation.

1. **TECHNICAL QUALITY**
   a. **Reliability**[3]
      i. *Study A*: Are business rules clear and complete?
      ii. *Study B*: Are scores consistent across years?
   b. Validity
      i. *Study C*: Are measures directly related to or strong proxies for student outcomes the system is intended to improve?
      ii. *Study D*: What unique information does each measure give to guide score use?
      iii. *Study E*: What schools are *consistent* outliers compared to their peer groups, and what are their relevant characteristics?
   c. Fairness
      i. *Study F*: Do measures relate as expected to demographics? Specifically, are good scores on non-achievement-based measures reasonably attainable for schools with low achievement?

2. **OUTREACH/COMMUNICATION**
   a. Communication
      i. *Study G*: Does WDE have a sound accountability communication plan (and was it implemented with fidelity)?
   b. Professional Learning
      i. *Study H*: Does WDE provide sound and relevant role-based professional learning resources?
      ii. *Study I:* Do stakeholders use learning resources at a reasonable rate (such as metrics obtained from web traffic and self-reports)?

3. **RECEPTION**
   a. Understanding
      i. Do school improvement teams, relevant district central office staff, and district policymakers accurately understand accountability reports and data as appropriate to their roles?
   b. Credibility
      i. Do school improvement teams, relevant district central office staff, and district policymakers see accountability reports and data as credible? Is this true of large, medium, and small size districts?

4. **UTILITY**
      i. Do school improvement teams, relevant district central office staff, and district policymakers appropriately use accountability reports and data to inform school improvement plans and to develop, evaluate, and refine policies and programs?

5. **EFFECTS**
      i. To what degree have teachers, school administrators, district central office staff, district policymakers, and state policymakers observed *unanticipated or*

[3] Traditional psychometric reliability analyses are not necessarily appropriate given that all scores other than input data (measures) are constructed variables (constructed on the basis of policy) rather than measures of some latent trait with multiple indicators/items.

unintended consequences of implementing the accountability system, with a particular focus on identifying negative unintended consequences?

ii. *Study J*: To what degree have *intended positive consequences* of improved student outcomes been empirically observed?

6. STABILITY

i. *Study K*: Is there a policy in place to guide potential revisions through a sound change management process incorporating (a) a principled evaluation of the rationale to determine if it is sufficiently important to merit disrupting the system, and (b) a principled approach to defining, developing, and incorporating the revision into the system in a manner that minimizes disruption and maximizes coherence with the rest of the system?

Level of Change

In addition to describing individual studies, it is important to determine when each study should be repeated in the event of a change to the system. Therefore, we label and define three levels of change to determine would trigger what kind of study repetition.

- *Major change*: A completely new system, addition of a new measure, removal of a measure, or regrouping of measures into indicators or indicators into indicator categories
- *Modest change*: A material change in score calculation (i.e., for measures, indicators, indicator categories, or overall school accountability ratings)
- *Minor change*: Any of the following:
  - Small changes to weights used in score calculations (e.g., less than a 10% change)
  - Technical edits to score calculation (e.g., correcting a small coding error)
  - Technical edits to business rule (e.g., such as correcting an ambiguous rule)

Prioritization

The Advisory Committee agreed that all of the studies outlined above might be important to conduct, but they recognized the importance of clearly prioritizing the studies to provide guidance for allocating limited resources to support the evaluation and continuous improvement work. Therefore, the committee participated in two prioritization activities. In the first, each member rank-ordered the research questions associated with each study. In the second, each member identified their top five studies. The results of these prioritization activities are shown in Appendix A. It was difficult to produce a fine-grained rank ordering for each study because of many close calls. Therefore, we present four levels of priorities below with priority #1 being the highest priority.

Priority 1

- Study C: Technical quality: Do the data measure what they are intended to measure?
- Study D: Technical quality: Is there a unique contribution of each measure?
- Study G: Outreach: Is there a sound communication plan implemented with fidelity?
- Study J: Effects: To what degree are intended consequences observed?

- Study I: Understanding: Do stakeholders understand accountability reports and data files?
    - Credibility: Do stakeholders find accountability reports credible?
    - Utility: Do stakeholders use reports/data as appropriate to their roles?
    - Effects: Do stakeholders see evidence of unintended consequences?

The estimated cost for all of the Priority 1 studies is approximately $250,000[4].

Priority 2

- Study F: Technical quality: Are non-achievement indicator targets attainable for all schools?

The estimated cost for the Priority 2 study is approximately $15,000.

Priority 3

- Study B: Technical quality: Are ratings consistent across years?

The estimated cost for the Priority 3 study is approximately $15,000.

Priority 4

- Study A: Technical quality: Business rules are clear and complete.
- Study E: Technical quality: Are there certain schools that are consistent outliers with respect to performance and associated characteristics?
- Study H: Professional learning: Are there sound learning resources available to stakeholders?
- Study K: Stability: Is there a sound policy in place to encourage long-term stability?

The estimated cost for all of the Priority 4 studies is approximately $137,000.

The estimated budget for each study is noted in the last column in Appendix A. The estimated budget numbers are drawn from the detailed descriptions of each study included in later sections. When there are multiple options for completing a study, the middle option (both in terms of rigor and cost) is included in Appendix A.

Potential Study Descriptions

Study A (Reliability): *Are Business Rules Adequately Clear and Complete?*

Approach 1: Critical Review of Business Rules for Clarity and Completeness

Business rule documents describe in detail how scores in the accountability system are calculated, including such minutiae as the details of data sources, data cleaning rules, inclusion and exclusion rules, and each step in the calculation process. This critical review should be

[4] Note: This does not mean that all of the studies will be conducted in a single year.

conducted by an analyst familiar with business rules. The rules should be critically reviewed for the following:

- Clarity (to address potential ambiguity in specifications for calculating accountability scores)
- Completeness (to document all assumptions about incoming data files and all steps in calculating accountability scores)

The level of effort associated with this type of review is moderate because business rules tend to be complex if they are adequate to allow for replication. This approach does not need to be repeated for subsequent years if there are no changes to the system. However, if there are *any* changes to business rules, it should be repeated in advance of releasing reports that incorporate the change.

Approach 2: Replication of Score Calculations

An important aspect of establishing the reliability, validity, and credibility of the system involves knowing that score calculations can be replicated by a competent analyst using the business rules used by the primary analyst(s) responsible for calculating scores. Replication substantially reduces but does not entirely eliminate the chance of reporting erroneous scores.

There are three basic methods for replication differing in level of rigor and level of effort (or resources/cost) required to conduct them as shown in Table 1. All methods require new code (intended to address potential coding errors). The methods differ in whether there is a different analyst and whether there are different software package used for replication than for official analyses. The less rigorous and costly approach is rated as such because it involves only one analyst in both the original and replication analyses. The other two require higher levels of effort because an additional analyst is involved. The most rigorous and costly option is rated as such because it involves two analysts and two software packages. Effort and cost in subsequent years is low because the original code/software package and replication code/software package may be reused without the need for recoding. Even for years with minor to modest changes, the original and replication code can be updated in a targeted manner without needing to write completely new code. Regardless of the option selected, scores should be replicated to within a defined level of tolerance to ensure that the same reported scores are created.

Table 1. *Approaches to replication.*

		Effort and Cost in…		Replication Uses Different…			Purpose is to Addresses Potential…		
Option	Rigor	Year 1	Years 2+	Code	Analyst	Software	Coding Errors	Ambiguous Business Rules	Software Bugs
A	Less	Modest	Low	✓		✓	✓		✓
B	More	High	Low	✓	✓		✓	✓	
C	Most	Highest	Low	✓	✓	✓	✓	✓	✓

It may be desirable to repeat the study every year even when there have been no changes to business rules as a safety check to maximize the likelihood of catching anomalies that slipped through the cracks the previous years because the data characteristics necessary to produce the anomalies were not present in previous years' data. However, this risk is smaller than the risk of bugs, differences across software packages, coding errors, and differences in understanding across coders. At a minimum, we recommend repeating this analysis in the second year (the mostly likely year to catch something that slipped through the cracks) and then every third year thereafter. Finally, replication should be conducted every time there is any change (even very minor changes) to business rules or operational code used for calculation. An estimated budget for initial implementation of study A for each option is provided in Appendix B.

Study B (Reliability): *Are Scores Reasonably Consistent across Years?*

Because the system will produce both continuous and categorical scores, monitoring needs to account for both types of scores, and because of consequences associated with aggregate scores, it is important to conduct these analyses for each level of score (*measure*, *indicator*, *indicator category*, and *overall*) separately. Because the *measure* score data are provided as inputs, the degree of consistency in data at the *measure* level can serve as a benchmark against which to compare the degree of consistency in data from the higher level (aggregate) scores. This includes benchmarking for small schools, which are of particular concern where stability is concerned.

For continuous data, correlational analyses are sufficient to evaluate the degree to which the ordering of schools on accountability scores changes from one year to another. For categorical data, contingency-table-based analyses are required. This is demonstrated in Figure 2 where yellow cells indicate consistent classifications and green and red cells mean inconsistent classifications with an improvement and a decline, respectively. Large numbers in the red and green cells are concerning, but numbers in the green cells should outweigh those in the red cells if the theory of action is correct.

Figure 2. *Sample contingency table for evaluating consistency of categorical scores over time.*

		Year 2 Category				
		Not Meeting Expectations	**Partially Meeting Expectations**	**Meeting Expectations**	**Exceeding Expectations**	**Row Total**
Year 1 Category	**Not Meeting Expectations**	**60%**	**28%**	**11%**	**1%**	**100%**
	Partially Meeting Expectations	**10%**	**57%**	**25%**	**8%**	**100%**
	Meeting Expectations	**3%**	**9%**	**78%**	**10%**	**100%**
	Exceeding Expectations	**0%**	**4%**	**11%**	**85%**	**100%**

TABLE NOTE: Cells shaded in yellow represent consistent classification from one year to the. Cells shaded in green (or red) represent inconsistent classification, with schools improving (or declining) from one year to the next.

The level of effort required to carry out this approach is modest because although there are many scores to analyze for consistency over time, the statistical analyses are simple. This study should be competed after the second round of reports is released (to allow for multiple years of data to evaluate consistency. We then recommend that it be repeated when modest to major changes are made to the system[5]. An estimated budget for the initial implementation of study B is provided in Appendix C.

Study C (Validity): *Are Measures Directly Related to or Strong Proxies for Student Outcomes the System is Intended to Improve?*

In this study, the following policy documents will be critically reviewed to identify the intended outcomes of and the intended meaning of components included in the accountability system:

- Authorizing and governing documents (legislation, policies, or rules)
- System design documents (recommended and adopted design documents)

Based on the understanding gained from the review of policy documents, business rules will be reviewed to evaluate alignment of the system in the following areas:

- The degree to which intended outcomes are represented in the accountability system.
- The degree to which data used by the system constitute measures of intended quantities.

This study should be completed as soon as possible after the first year of reporting so that it can be used to inform the first set of evaluations. This analysis does not need to be repeated in subsequent years with two exceptions. If a major or modest change is made, a full analysis should be repeated in advance of completion of revised code. If a minor change is made, a partial (incremental) analysis should be conducted related to the specific changes made. The level of effort associated with this study is moderate as there may be a number of documents needing review and because business rules and some policy documents tend to be complex. An estimated budget for the initial implementation of study C is provided in Appendix D.

Study D (Validity): *What Unique Information does Each Measure Give to Guide Score Use?*

This study will use basic multiple regression analyses to estimate how much information each measure score contributes to each higher-level aggregate score[6]. This approach can be implemented by standardizing all measure scores and aggregate scores and running a multiple regression of each aggregate score on all of the standardized contributing measure scores. The resulting standardized regression coefficients represent the relative contributions of each measure

[5] When minor changes are made, it is unnecessary to repeat these analyses as minor changes tend to affect a small number of school scores and would very likely have small effects on statewide stability of scores.

[6] That is, indicator scores, indicator category scores, and overall scores to which more than one measure contributes.

score to the aggregate score[7], as compared to the weights assigned to each measure in constructing the aggregate score.

This approach should be used for all measures that contribute to each indicator, all measures that contribute to each indicator category, and all measures that contribute to overall scores to identify which measures have the greatest effective weights in the constructed variables to which they contribute. This study requires moderate effort as the statistical analyses are relatively simple but become potentially complicated because of missing data. It should be conducted after the first cycle of accountability reporting and after any modest to large change to the system. An estimated budget for the initial implementation of study D is provided in Appendix E.

Study E (Validity): *What Schools are Consistent Outliers Compared to their Peer Groups, and What are their Relevant Characteristics?*

Step 1: Peer Group Identification and School Location within the Group

Based on the way the accountability system is structured, it is most reasonable to identify consistent outlier schools on the basis of measure scores. A generic description of approaches to identifying consistent outlier schools is that moderately sophisticated statistical analyses are conducted to perform the following steps for each measure[8]:

1. Identify a *hypothetical* or *actual* peer group for each school in each year
2. Determine whether each school is a low or a high outlier within its peer group each year
3. Determine whether outlier identification is consistent over some number of years

We recommend two types of analyses to accomplish these steps. The first type is based on identifying a *hypothetical* peer group based on statistically predicted scores of hypothetical schools with the same demographics. It can be carried out using simple multiple regression of indicator scores on school demographics to evaluate the difference between predicted and observed measure scores. Schools would be identified as outliers for a given year if their predicted and observed scores are statistically significantly different from each other at some reasonable α-level. The second type of analysis is based on identifying for each school an *actual* peer group with the most similar school demographics. A procedure for identifying this peer group is described in the Appendix F. Schools would be identified as *consistent* outliers if they are identified as outliers using either method for three consecutive years to ensure that effort expended in the next approach is based on truly consistent outliers. The effort level required for this approach is moderate in that it requires the use of some moderately sophisticated statistical methods to identify peer groups and school location within the group. This analysis should be conducted after the third cycle of accountability reporting to allow for identifying *consistent* outlier schools.

[7] This procedure calculates standardized β coefficients in a multiple regression, a measure of effect size in predicting a dependent variable. Because the procedure creates a saturated model perfectly predicting the constructed variable in most cases (and nearly perfectly in the remaining cases), the results can be interpreted as relative contributions.

[8] An analysis should be conducted for each indicator because different school characteristics are likely to be more or less relevant to each indicator.

Step 2: Engagement with Stakeholders from Outlier Schools

Schools identified as consistent outliers on a given indicator can then be engaged using one of the three methods of stakeholder engagement (described in the section on study I) to gather information about the schools' practices, the way the school is structured, and the nature of the school's relationship with the district. If feasible, we recommend that this be conducted separately for two groups of outlier schools: (1) those that outperform their peer group, and (2) those that underperform their peer group to allow for contrasting the characteristics of these two groups. If this is not feasible, it is reasonable to focus only on those that outperform their peer groups.

We recommend that for this approach, targeted phone interviews be conducted at a minimum to avoid issues with low response rate. The purpose of these activities is to determine whether well-implemented educational practices, structures, and productive relationships between school and district identified in research literature as associated with better performance on test-based measures are also associated with measures used in the accountability system. Finally, this approach might also be used to identify potential best practices to share with all schools.

This step should be implemented after step 2 is complete, and then periodically (e.g., every 3-5 years) thereafter to allow for initial identification of consistent outliers, and then to continue to build on validity evidence and a potential stable of best practices. The overall level of effort needed for this approach is very high as it involves four high-effort steps:

- Reviewing the literature on practices, structures, and relationships that are effective in improving student outcomes
- Developing surveys, interview protocols, or focus group protocols
- Conducting surveys, interviews, or focus groups
- Cleaning, coding, and analyzing collected data in light of the literature review

An estimated budget for initial implementation of study E for each option is provided in Appendix G.

Study F (Fairness)

It is important that the accountability system not unreasonably assign lower or higher scores to schools based on the student populations they serve.

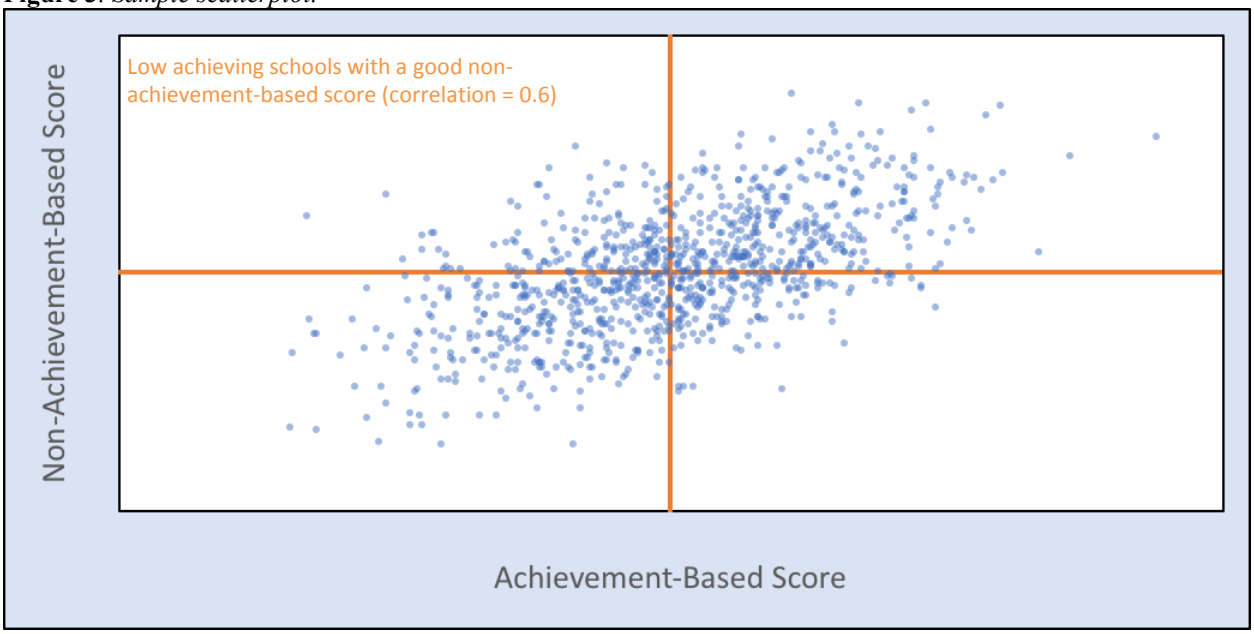
Question 1: *Do Measures Relate as Expected to Demographics?*

It is well known that achievement scores are strongly related to many demographic variables. Itis reasonable to expect that will continue to be the case for achievement-based measures such as those that comprise the *achievement* indicator and to a lesser degree those that comprise the *post-secondary readiness* indicator. However, relationships with demographics should be much reduced for the remaining indicators. This analysis can be conducted by comparing the percent of variation explained by demographics for achievement-based measures vs. other measures. Based on reviews of the literature and logical evaluations of whether schools have reasonable approaches to ameliorate demographic differences in other measures, one might expect to see weaker relationships with measures that contribute to the *growth*, *equity*, *English language progress*, *graduation rate*, and *grade-9 credits* indicators.

Question 2: *Are Good Scores on Non-Achievement-Based Measures Reasonably Attainable for Schools with Low Achievement Scores?*

This analysis can be conducted by evaluating the strength of relationships between achievement-based measures and non-achievement-based measures. For relationships between continuous scores, this can be done statistically using correlations to identify relationships that are strong enough to cause concern, or graphically as demonstrated in Figure 3 where a "good score" is considered to be above average and "low achievement" is considered to be below average. These definitions could be altered to represent the "met expectations" cut score set by the professional judgment panel (PJP).

Figure 3. *Sample scatterplot.*



For relationships between categorical scores, this can be done graphically as demonstrated in Figure 4, where "good" non-achievement based scores are defined as higher than achievement-based scores (the numbers in the green cells represent how common these "good" scores are).

Figure 4. *Sample contingency table.*

		Non-Achievement-Based Score				
		Not Meeting Expectations	Partially Meeting Expectations	Meeting Expectations	Exceeding Expectations	Row Total
Achievement-Based Score	Not Meeting Expectations	60%	28%	11%	1%	100%
	Partially Meeting Expectations	10%	57%	25%	8%	100%
	Meeting Expectations	3%	9%	78%	10%	100%
	Exceeding Expectations	0%	4%	11%	85%	100%

In either case (continuous or categorical data), a first step will be to define the following terms:

- A good non-achievement-based scores
- Low achievement
- Reasonably attainable

These analyses should be conducted after the first year of reporting, and then every third year thereafter only when a modest or major change is implemented. An estimated budget for the initial implementation of study F is provided in Appendix H.

Study G (Communication): *Does WDE have a sound accountability communication plan (and was it implemented with fidelity)?*

This study will be based on a critical review of WDE's plan for communicating with stakeholders regarding accountability *reports*, *data*, and *professional learning resources* for stakeholders to gain understanding of the reports and how to use them in ways consistent with their role (e.g., in making decisions about school improvement and policy or program development, evaluation, and refinement). It will also be based on a critical review of actual communications with the field to evaluate whether the plan was carried out as designed (with reasonable modifications to address issues as needed). The primary purpose of this study is to inform potential improvements to the communication plan and its implementation.

The level of effort for this study is moderate in that a communication plan that is effective in bringing learning resources to the attention of stakeholders is likely to be modestly complex, and the number of communications to be reviewed may be modestly large. This study should be conducted by an analyst with expertise in developing and implementing communication plans after the first cycle of accountability reporting and should be followed up the next year if

substantial recommended improvements are identified. An estimated budget for the initial implementation of study G is shown in Appendix I.

Study H (Professional Learning): *Does WDE provide sound and relevant role-based professional learning resources?*

This study is based on a critical review of learning resources provided for stakeholders who use accountability reports and data in their roles. The specific uses of reports and data are the intended uses of (a) school improvement planning and implementation, and (b) developing, evaluating, and refining local policies and programs. The critical review should address each of the following elements:

- Sound instructional design
- Sound instruction for understanding accountability reports accurately, clearly, and unambiguously
- Resources demonstrating appropriate use of accountability reports and data that are relevant and compelling to stakeholders

The groups of stakeholders for which the demonstrated uses should be relevant and compelling include teachers, school administrators, district central office staff, district administrators, and possibly school board members responsible for the following:

- School improvement planning
- School program development/evaluation/refinement
- School policy development/evaluation/refinement
- District program development/evaluation/refinement
- District policy development/evaluation/refinement

The purpose of the critical review is to inform potential improvements to the professional learning resources. The resources needed for this study are high because the potential number of resources is large, and they may be complex. This study should be conducted after the first cycle of accountability reporting by an analyst familiar with both the accountability system and instructional design. It may need to be followed up in subsequent years if material improvements to the learning resources are identified in the first year or improvements to the communication plan are made to improve market penetration. An estimated budget for initial implementation of study H is provided in Appendix J.

Study I: *Learning Resources, Understanding, Credibility, and Utility*

In this study, several research questions will be addressed because they all require direct engagement with stakeholders and they all should be conducted at the same time. The research questions are:

- Do stakeholders use learning resources at a reasonable rate (from self-reports)?

- Do school improvement teams, relevant district central office staff and district policymakers accurately understand accountability reports and data as appropriate to their roles?
- Do school improvement teams, relevant district central office staff, and district policymakers see accountability reports and data as credible? Is this true of both large, medium, and small size districts?
- Do school improvement teams, relevant district central office staff, and district policymakers appropriately use accountability reports and data to inform school improvement plans and to develop, evaluate, and refine policies and programs?
- To what degree have teachers, school administrators, district central office staff, district policymakers, and state policymakers observed[9] *unanticipated or unintended consequences* of implementing the accountability system (with a particular focus on negative consequences[10])?

All of these questions should be answered for multiple stakeholders as theorized in the theory of action to drive improvements in student outcomes. Those groups are stakeholders responsible for the following[11]:

- School improvement planning
- School program development/evaluation/refinement
- School policy development/evaluation/refinement
- District program development/evaluation/refinement
- District policy development/evaluation/refinement

Answering these questions requires a concerted effort to engage directly with stakeholders to gather the necessary data. This can be conducted in one of three ways as shown in Table 2. Targeted follow-up (calls or emails) are included for all approaches because an interesting or ambiguous comment or phenomenon may be identified during analysis that merits following up.

Table 2. *Approaches to stakeholder engagement for the purpose of evaluating the accountability system.*

			Effort/Cost in…	
Approach (each is based on representative sampling)		**Rigor**	**First Year**	**Subsequent Years**
Online surveys	+ targeted follow-up	Less	Modest	Low
Targeted phone interviews	+ targeted follow-up	More	High	Modest
Focus groups	+ targeted follow-up	Most	High	Modest

[9] The surveys, interview protocols, or focus group protocols should distinguish between rigorous evidence (i.e., intentionally gathered an analyzed evidenced) and anecdotal evidence. For what is likely to be the majority (anecdotal evidence) respondents should be asked to identify the source of evidence if it is feasible for them to do so (so patterns of similar sources giving similar anecdotal evidence can be identified).

[10] Such as the following: (a) changes to how students tend to be routed into high school courses based on perceived or real differences in difficulty of meeting the three options for credit in the *postsecondary readiness* indicator and (b) reduced attention to subjects other than English language arts or mathematics.

[11] Drawn from teachers, school administrators, district central office, district administrators, and possibly school board members.

Surveys tend to have low response rates, but they also tend to require less effort and cost. It might be possible to improve the quality of data gathered through surveys by developing a plan to sample stakeholders, with personal phone calls from a feasibly high-level WDE official to encourage sampled stakeholders to respond to the surveys. In addition, there is a lot going on with this study, and a number of surveys would need to be targeted to representative samples of stakeholders so that no survey is so lengthy as to cause a low completion rate.

Targeted phone interviews with representative samples of relevant stakeholders are likely to increase response rate as they indicate a greater investment on the part of WDE, but they require increased effort and cost. Similar to the issue with survey length, because of the number of questions to be addressed in this study, a number of samples would need to be drawn to keep each individual phone interview to a reasonable length.

Focus groups may be the most appropriate option because it should be possible to address the complete set of questions in one-day focus groups conducted in a small number of locations around the state with carefully selected focus group members. While costs of focus groups may include facilitator travel to the focus group sites, it may be possible to eliminate the need for large travel reimbursements by inviting panelists from nearby the site. The costs for meeting facilities might also be minimized by convening focus groups on school or district properties. A final reason that focus groups may be a better option is stakeholders will have the benefit of listening to each other in responding, producing more thoughtful responses and more useful data. This study should be conducted sufficiently after the accountability reporting cycle to allow time for stakeholder perceptions to mature somewhat, for stakeholders to use the professional learning resources, and for stakeholders to use accountability reports/data for school improvement or program/policy development, evaluation, and refinement decisions. We also recommend that these be conducted after the second year and every 3-5 years after that. We recommend ongoing data collection and analysis to (a) identify potential improvements to learning resources and associated communications, and (b) identify issues with credibility and what might be done to address then as such issues arise. An estimated budget for the initial implementation of study I is provided in Appendix K.

Study J: *To what degree have intended positive consequences (including improved student outcomes) been empirically observed?*

The ultimate question is whether the theory of action has been realized. This study should use longitudinal modeling (i.e., student growth) of the measures used in the accountability system to determine whether there is any evidence of improvement in student outcomes as intended[12]. The Wyoming Test of Proficiency and Progress (WY-TOPP) is new so the availability of baseline data for measures in the *academic performance* indicator category and the *equity* indicator may be restricted to a single year. However, it may be possible to collect longer-term baseline data for the *English language progress* indicator and the measures in the *readiness* indicator category for high school. This study should be conducted using multilevel statistical models that nest time series data[13] within schools to accomplish the following (a) account for multiple data points coming from the same school, (b) account for potential differences in rate of change in student

[12] Establishing that implementation of the system cause the improvements is a higher bar.

[13] Or data from a series of years.

outcomes across schools, and (c) to model differences in rate of change using school demographics.

Results of this study can reasonably be used to identify potential areas in the long term outcomes have been realized, but will be insufficient to *definitively* establish that implementing the system caused the changes in student outcomes. However, if improvements are sustained and/or accelerated over multiple years, attributing such improvements to implementing the system (with the associated elements of the theory of action) will become a more reasonable proposition.

The effort required for this study is high in that it requires a sophisticated analytical approach, addresses a considerable number of outcomes[14], and incorporates a number of independent variables (school demographics) in addition to the passage of time. However, given that existing code for running the multilevel models can be used in subsequent years, subsequent analyses will require considerably less effort. This study should first be conducted after the fourth year[15] of score reporting and repeated periodically thereafter (e.g., every 3-5 years). In addition, years in which changes to the system are made might be incorporated into the models as predictors or rate of change to identify potential effects of changes to the system. An estimated budget for the initial implementation of study J is provided in Appendix L.

Study K: *System Stability as a Necessary Condition for Realizing the Theory of Action*

This study answers the following question: *Is there a policy in place to guide potential revisions through a sound change management process incorporating (a) a principled evaluation of the rationale to determine if it is sufficiently important to merit disrupting the system, and (b) a principled approach to defining, developing, and incorporating the revision into the system in a manner that minimizes disruption and maximizes coherence with the rest of the system?*
This analysis should be conducted using any policy documents that address system stability over time, critically evaluating those documents based on whether the policy, if followed, would maximize the likelihood of (a) achieving adequate system stability over time and (b) maintaining a coherent design.

This analysis should be conducted after the second year of implementation to allow time for a change management process to be developed and adopted, whether adoption is through legislation, rule, or policy. It should be repeated after any major change to the system. The effort required for this study is low as it only requires critical review of a small potential number of relevant policy documents. An estimated budget for the initial implementation of study K is provided in Appendix M.

[14] Eight elementary and middle schools (EMS) measures, 13 high schools (HS) measures, 4 EMS indicators, and 7 HS indicators.

[15] To ensure that growth rates are based on more than just two years of data. It can be demonstrated mathematically that the slope of a linear regression through three equally-separated data points (such as those gathered on an annual basis) depends solely on the first and third data point (the intercept depends on all three data points, but the intercept is not useful for answering the evaluation question).

REFERENCES

D'Brot, J., Lyons, S., & Landl, E. (2017). *State Systems of Identification and Support under ESSA: Evaluating Identification Methods and Results in an Accountability System*. Washington, DC: Council of Chief State School Officers. Retrieved July 15, 2018. Available at https://www.nciea.org/library/state-systems-identification-and-support-under-essa-evaluating-identification-methods-and.

Appendix A: Study Descriptions, Estimated Costs, and Prioritization

Study	Brief Study Description	Priority	Rank	#Votes	Estimated Cost
C	Technical quality data measure what they are intended to measure	1	5	10	$6,650
D	Technical quality unique contribution of each measure	1	5	13	$14,250
G	Outreach is there a sound communication plan implemented with fidelity?	1	4	4	$37,800
K	Effects to what degree are intended consequences observed?	1	9	9	$92,300
J	Professional learning: Do stakeholders use the learning resources (self-report)?	1	12	12	$91,350
	Understanding: Do stakeholders understand accountability reports and data files?				
	Credibility: Do stakeholders find accountability reports credible?				
	Utility: Do stakeholders use reports/data as appropriate to their roles?				
	Effects: Do stakeholders see evidence of unintended consequences?				
F	Technical quality: Are non-achievement indicator targets attainable for all schools?	2	8	8	$15,200
B	Technical quality: Are scores consistent across years	3	6	6	$14,250
A	Technical quality: Are business rules are clear and complete?	4	10	10	$22,800
E	Are there certain schools that are consistent outliers with respect to performance and associated characteristics?	4	11	11	$80,900
I	Professional learning: Are there sound learning resources available to stakeholders?	4	11	11	$29,800
L	Stability: Is there a sound policy in place to encourage long-term stability?	4	12	12	$2,400

Appendix B: Estimated budget for study A.

		Estimated					
		Option A		Option B		Option C	
Approach	**Major Activity**	**Effort**	**Cost**	**Effort**	**Cost**	**Effort**	**Cost**
Critical Review of Business Rules	Critical review of business rules for clarity and completeness	1.5 days	$2,850	1.5 days	$2,850	1.5 days	$2,850
	Deliver first draft of annotated business rules identifying any lack of clarity and completeness with associated recommendations for improvement	0.5 days	$950	0.5 days	$950	0.5 days	$950
	Meet with WDE staff and consultants responsible for business rules to discuss first draft and associated recommendations	0.5 days	$950	0.5 days	$950	0.5 days	$950
	Deliver final draft incorporating WDE staff feedback as appropriate	0.5 days	$950	0.5 days	$950	0.5 days	$950
Replication	1st analyst writing code in 2nd software package, checking agreement with results from 1st analyst's code in 1st software package, and resolving the discrepancies until sufficient agreement is reached	5.0 days	$9,500	–	–	–	–
	2nd analyst receiving, reviewing, and cleaning data; writing code in 1st software package, checking agreement with results from code written by 1st analyst in 1st software package, collaborating to resolve discrepancies until sufficient agreement is reached	–	–	9.0 days	$17,100	–	–
	2nd analyst receiving, reviewing, and cleaning data; writing code in 2nd software package, checking agreement with results from code written by 1st analyst in 1st software package, and collaborating to resolve discrepancies until sufficient agreement is reached	–	–	–	–	11.0 days	$20,900
	Total		**$15,200**		**$22,800**		**$26,600**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix C: Estimated budget for study B.

Major Activity	Estimated Effort	Estimated Cost
Collaboration on data exchange and extraction	1.0 days	$1,900
Data cleaning	1.0 days	$1,900
Data analysis	2.0 days	$3,800
Deliver a first draft of report on year to year consistency to WDE and any associated recommendations	2.0 days	$3,800
Deliver a second draft to WDE and the Advisory Committee (attending to WDE feedback on the first draft report as appropriate)	1.0 days	$1,900
Deliver a final draft to WDE, the State Board of Education, and the Accountability Advisory Committee (attending to feedback on the second draft as appropriate)	0.5 days	$950
	Total	**$14,250**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix D: Estimated budget for study C.

Major Activity	Estimated Effort	Estimated Cost
Gathering governing documents and system design documents,	1.0 days	$1,900
Evaluating the match between governing documents and system design documents and whether measures used represent direct or close proxies for intended quantities	1.0 days	$1,900
Deliver a first draft of report on consistency between governing documents and system design documents to WDE	0.5 days	$3,800
Deliver a second draft to WDE and the Advisory Committee (attending to WDE feedback on the first draft report as appropriate)	0.5 days	$1,900
Deliver a final draft to WDE, the State Board of Education, and the Accountability Advisory Committee (attending to feedback on the second draft as appropriate)	0.5 days	$950
	Total	**$6,650**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix E: Estimated budget for study D.

Major Activity	Estimated Effort	Estimated Cost
Collaboration on data exchange and extraction	1.0 days	$1,900
Data cleaning	1.0 days	$1,900
Data analysis (including potential multiple imputation)	2.0 days	$3,800
Deliver a first draft of report on relative contributions (or effective weights) to WDE	2.0 days	$3,800
Deliver a second draft to WDE and the Advisory Committee (attending to WDE feedback on the first draft report as appropriate)	1.0 days	$1,900
Deliver a final draft to WDE, the State Board of Education, and the Accountability Advisory Committee (attending to feedback on the second draft as appropriate)	0.5 days	$950
	Total	**$14,250**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix F: Example Methodology for Identifying *Actual* Peer Groups for Each School

It can be carried out by creating a vector of demographic data for each school with at a minimum the following data:

- Number of students enrolled
- Location in a rural area (coded 0/1)
- Percent in each race/ethnicity category (with statewide sample sizes sufficient for use)
- Percent of students with disabilities
- Percent of students that are English learners
- Percent of students that are economically disadvantaged

The last four might be replaced by the percent of students that are in the consolidated subgroup. The data in the vectors are normalized (to a mean of zero, variance of 1) to weight each demographic variable equally. The Euclidean distance between every pair of schools with a score[16] is then calculated as:

$$D = \sqrt{\sum_{k=1}^{K} \left(v_{ik} - v_{jk}\right)^2}$$

Where K is the number of demographic variables, k indicates a specific demographic variable, v_{ik} indicates the value of demographic variable k for school i, and v_{ik} indicates the value of that variable for school j. A peer group for each school would then be identified by finding the N schools with the smallest distances from the school the peer group is for. Schools would be identified as outliers for a given year if they are the top or bottom school in their peer group (or in the top 2 or bottom 2 if larger peer groups are formed).

[16] That is, every pair of schools with an elementary/middle school score and every pair of schools with a high school score.

Appendix G. Estimated budget for study E.

		Estimated								
		Survey option			Phone interview option			Focus group option		
Step	**Major Activity**	**Expenses**	**Effort**	**Cost**	**Expenses**	**Effort**	**Cost**	**Expenses**	**Effort**	**Cost**
Identify Outlier Schools	Collaboration on data exchange and extraction	–	1.5 days	$2,850	–	1.5 days	$2,850	–	1.5 days	$2,850
	Data cleaning	–	0.5 days	$950	–	0.5 days	$950	–	0.5 days	$950
	Data analysis	–	3.0 days	$5,700	–	3.0 days	$5,700	–	3.0 days	$5,700
	Deliver 1st draft of anonymized report about identified outlier schools to WDE	–	1.0 days	$1,900	–	1.0 days	$1,900	–	1.0 days	$1,900
	Deliver 2nd draft of anonymized report (incorporating WDE feedback as appropriate)	–	0.5 days	$950	–	0.5 days	$950	–	0.5 days	$950
Engage stakeholders from select outlier schools to understand their characteristics	Develop a rationale with WDE for strategic sampling of outlier schools for engagement	–	5.0 days	$9,500	–	5.0 days	$9,500	–	5.0 days	$9,500
	Literature review of practices, structures, & relationships effective in improving student outcomes	–	7.0 days	$13,300	–	7.0 days	$13,300	–	7.0 days	$13,300
	Develop surveys	–	4.0 days	$7,600	–	–	–	–	–	–
	Develop interview protocols	–	–	–	–	5.0 days	$9,500	–	–	–
	Develop focus group protocols	–	–	–	–	–	–	–	5.0 days	$11,400
	Conduct surveys	–	1.0 days	$1,900	–	–	–	–	–	–
	Conduct interviews	–	–	–	–	4.0 days	$7,600	–	–	–

Step	Major Activity	Estimated								
		Survey option			Phone interview option			Focus group option		
		Expenses	Effort	Cost	Expenses	Effort	Cost	Expenses	Effort	Cost
	Conduct focus groups (4 focus groups)	–	–	–	–	–	–	$6,000	6.0 days	$11,400
	Clean and code data	–	1.0 days	$1,900	–	2.0 days	$3,800	–	3.0 days	$7,600
	Analyze data	–	2.0 days	$3,800	–	2.0 days	$3,800	–	3.0 days	$5,700
	Deliver a 1st draft of a report on characteristics of low and high outlier schools to WDE	–	1.5 days	$2,850	–	2.0 days	$3,800	–	3.0 days	$5,700
	Deliver a 2nd draft to WDE and Advisory Committee (with feedback from WDE as appropriate)	–	1.0 days	$1,900	–	1.0 days	$1,900	–	2.0 days	$3,800
	Deliver a final draft to WDE, Advisory Committee, and State Board of Education	–	0.5 days	$950	–	0.5 days	$950	–	1.0 days	$1,900
	Optional: Testify before the State Board of Education	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200
	Optional: Testify before the appropriate (Joint) Legislative Committee	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200
	Total without optional line items		**29.5 days**	**$56,050**		**35.0 days**	**$66,500**		**41.5 days**	**$84,850**
	Total with optional line items		**35.5 days**	**$70,450**		**41.0 days**	**$80,900**		**47.5 days**	**$99,250**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix H. Estimated budget for study F.

Major Activity	Estimated Effort	Estimated Cost
Defining terms in collaboration with WDE	0.5 days	$950
Collaboration on data exchange and extraction	1.0 days	$1,900
Data cleaning	1.0 days	$1,900
Data analysis	2.0 days	$3,800
Deliver 1st draft of report on fairness to WDE	2.0 days	$3,800
Deliver 2nd draft to WDE and the Advisory Committee (attending to WDE feedback as appropriate)	1.0 days	$1,900
Deliver final draft to WDE, the State Board of Education, and the Advisory Committee (attending to feedback on the 2nd draft as appropriate)	0.5 days	$950
	Total	**$15,200**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix I: Estimated budget for study G.

Major Activity	Estimated Expenses	Effort	Cost
Become familiar with key audiences, accountability reports, accountability data files, and accountability professional learning resources	–	5.0 days	$6,000
Critically review the communications plan as a strategic approach to maximizing audience awareness, understanding, and use of accountability reports, data files, and professional learning resources	–	3.0 days	$3,600
Critically review the implementation of the communications plan	–	3.0 days	$3,600
Deliver a first draft of report of the critical review with recommendations for enhancements for a year-2 communications plan.	–	3.0 days	$3,600
Deliver a second draft to WDE and the Advisory Committee (attending to WDE feedback on the first draft report as appropriate)	–	2.0 days	$2,400
Attend an in-person meeting with the Advisory Committee to discuss and take feedback on the second draft.	$1,500	3.0 days	$5,100
Deliver a final draft to WDE, the State Board of Education, and the Accountability Advisory Committee (attending to feedback on the second draft as appropriate)	–	1.0 days	$1,200
Optional: Testify before the State Board of Education	$1,500	3.0 days	$5,100
Optional: Testify before the appropriate (Joint) Legislative Committee	$1,500	3.0 days	$5,100
		Total without optional costs	**$27,600**
		Total with optional costs	**$37,800**

NOTE: Expense rate is calculated based on a round-trip flight from out of state into Casper, two nights' lodging, and three days' meals, transportation, and incidentals. Daily rate estimated based on the October 5, 2018 indeed.com reported median annual salary for a Director of Communications at $88,540 * 2 / 260 = $681/day rounded up to $700/day. A premium of $500/day is added for experience with strategic communication regarding complex calculations, data, and reports to non-technical stakeholders for **$1,200/day**.

Appendix J: Estimated budget for study H.

Major Activity	Estimated Expenses	Effort	Cost
Become familiar with key stakeholder roles, accountability reports, and accountability data files	–	3.0 days	$3,300
Critically review the professional learning resources as accurate, instructionally sound, well-designed, engaging, and relevant to key stakeholder roles	–	5.0 days	$5,500
Deliver a first draft of report of the critical review with recommendations for enhancements for a year-2 communications plan.	–	3.0 days	$3,300
Deliver a second draft to WDE and the Advisory Committee (attending to WDE feedback on the first draft report as appropriate)	–	2.0 days	$2,200
Attend an in-person meeting with the Advisory Committee to discuss and take feedback on the second draft.	$1,500	3.0 days	$4,800
Deliver a final draft to WDE, the State Board of Education, and the Accountability Advisory Committee (attending to feedback on the second draft as appropriate)	–	1.0 days	$1,100
Optional: Testify before the State Board of Education	$1,500	3.0 days	$4,800
Optional: Testify before the appropriate (Joint) Legislative Committee	$1,500	3.0 days	$4,800
		Total without optional costs	**$20,200**
		Total with optional costs	**$29,800**

NOTE: Expense rate is calculated based on a round-trip flight from out of state into Casper, two nights' lodging, and three days' meals, transportation, and incidentals. Daily rate estimated based on the October 5, 2018 indeed.com reported median annual salary for an e-Learning Manager at $75,647 * 2 / 260 = $581/day rounded up to $600/day. A premium of $500/day is added for experience with managing non-technical stakeholder e-learning incorporating complex calculations, data, and reports for **$1,100/day**.

Appendix K: Estimated budget for study I.

Major Activity	Estimated								
	Survey option			Phone interview option			Focus group option		
	Expenses	Effort	Cost	Expenses	Effort	Cost	Expenses	Effort	Cost
Develop a rationale with WDE for strategic sampling of schools for engagement	–	1.0 days	$1,900	–	1.0 days	$1,900	–	1.0 days	$1,900
Develop surveys	–	6.0 days	$11,400	–	–	–	–	–	–
Develop interview protocols	–	–	–	–	8.0 days	$15,200	–	–	–
Develop focus group protocols	–	–	–	–	–	–	–	8.0 days	$15,200
Conduct surveys	–	1.0 days	$1,900	–	–	–	–	–	–
Conduct interviews	–	–	–	–	10.0 days	$19,000	–	–	–
Conduct focus groups (8 focus groups)	–	–	–	–	–	–	$8,000	10.0 days	$19,000
Clean and code data	–	2.0 days	$3,800	–	4.0 days	$7,600	–	6.0 days	$11,400
Analyze data	–	5.0 days	$9,500	–	7.0 days	$13,300	–	9.0 days	$17,100
Deliver a 1st draft of a report to WDE	–	4.0 days	$7,600	–	6.0 days	$11,400	–	8.0 days	$15,200
Deliver a 2nd draft to WDE and Advisory Committee (with feedback from WDE as appropriate)	–	2.0 days	$3,800	–	3.0 days	$5,700	–	3.0 days	$5,700
Deliver a final draft to WDE, Advisory Committee, and State Board of Education	–	1.0 days	$1,900	–	1.5d	$2,850	–	1.5d	$2,850
Optional: Testify before the State Board of Education	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200
Optional: Testify before the appropriate (Joint) Legislative Committee	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200	$1,500	3.0 days	$7,200

Major Activity	Estimated								
	Survey option			Phone interview option			Focus group option		
	Expenses	Effort	Cost	Expenses	Effort	Cost	Expenses	Effort	Cost
Total without optional line items			**$41,800**			**$76,950**			**$89,650**
Total with optional line items			**$56,200**			**$91,350**			**$104,050**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix L: Estimated budget for study J.

Major Activity	Estimated Expenses	Effort	Cost
Become familiar with the accountability system, available data elements, and engaging in requirements gathering with WDE for data exchange and general analyses to be conducted	–	5.0 days	$9,500
Collaboration on data exchange and extraction	–	3.0 days	$5,700
Data cleaning	–	3.0 days	$5,700
Develop quasi-causal methodology for evaluating the effects of the system	–	10.0 days	$19,000
Model development and finalization	–	10.0 days	$19,000
Deliver 1st draft of report to WDE	–	6.0 days	$11,400
Deliver 2nd draft to WDE and the Advisory Committee (based on 1st draft feedback)	–	3.0 days	$5,700
Deliver a final draft to WDE, the State Board of Education, and the Accountability Advisory Committee (based on 2nd draft feedback)	–	1.0 days	$1,900
Optional: Testify before the State Board of Education	$1,500	3.0 days	$5,700
Optional: Testify before the appropriate (Joint) Legislative Committee	$1,500	3.0 days	$5,700
		Total without optional costs	**$80,900**
		Total with optional costs	**$92,300**

NOTE: Daily rate estimated at $1,900/day for performing specialized assessment and accountability related data analysis and documentation services. No job classification on indeed.com with the breadth of experience required to perform this work in accountability without a significant and costly learning curve.

Appendix M. Estimated budget for study K.

Major Activity	Estimated Expenses	Effort	Cost
Gathering governing documents and policy documents in collaboration with WDE and discussing the existing safeguards (if any) with WDE officials	$1,500	2.0 days	$91,200
Evaluating the degree to which there is a policy either in statute or rule that discourages changing the system (including the assessment system) without a highly compelling reason, and only then in a deliberate and thoughtful manner that maintains as much consistency as possible	–	1.0 days	$,600
Deliver a draft statement about the adequacy of the policy and potential recommendations for improvement to WDE, the State Board of Education, and the Accountability Advisory Committee	–	1.0 days	$600
		Total	**$2,400**

NOTE: Expense rate is calculated based on a round-trip flight from out of state into Casper, two nights' lodging, and three days' meals, transportation, and incidentals. Daily rate estimated based on the October 5, 2018 indeed.com reported median annual salary for a Senior Policy Advisor at $76,314 * 2 / 260 = $587/day rounded up to $600/day.