A Framework to Support Accountability Evaluation

Erika Hall and Chris Domaleski, National Center for the Improvement of Educational Assessment

Mike Russell, Boston College

Laura Pinsonneault, Wisconsin Department of Public Instruction

Prepared for the CCSSO ASR SCASS

Introduction

Education accountability systems are designed to improve student achievement by providing information that supports state leaders, districts and schools in making informed decisions about educational quality, teaching and learning. To examine whether they are working as intended and inform system improvements, thoughtful evaluation of the design and implementation of these systems is necessary. The purpose of this paper is to: a) present a framework that identifies the key elements underlying a comprehensive accountability system evaluation, and b) illustrate, at a high level, how this framework can be applied to inform the design of evaluation plans for state and locally defined education accountability systems.

Introduction to the Framework

Education accountability systems have many moving and inter-dependent parts. Evaluating the effectiveness of an educational accountability system requires a solid understanding of each of these parts, the role they are intended to play within the system, the short-term effects they are intended to have, and the combined overall effects the system is designed to have. The framework presented here is intended to help programs design an informative evaluation that provides insight into the extent to which specific parts of the system are implemented with fidelity, are having their intended short-term effects, and whether the system as a whole is working as designed to produce the intended outcomes.

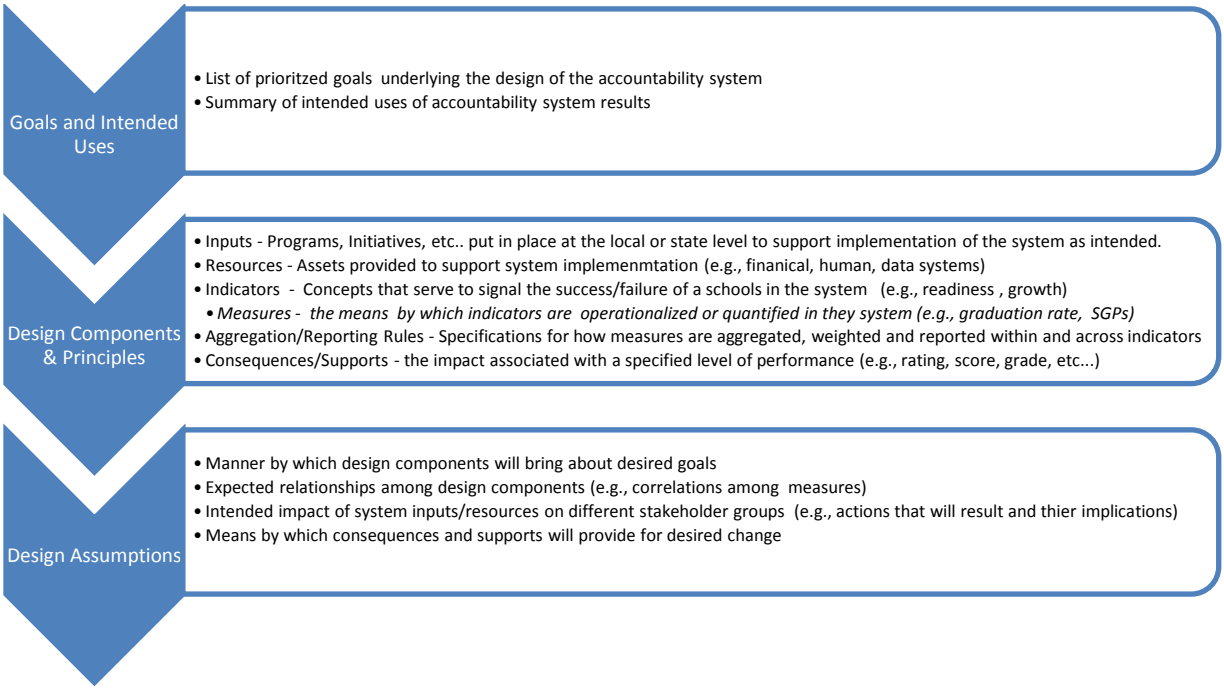
A key tenet underpinning the framework is that to be effective, an accountability system must be coherent. In a coherent system, the state's goals are clearly defined and the indicators of success are identified, operationalized, and emphasized in a manner consistent with the state's goals. In addition, the inputs and resources provided by the state or by local entities are thoughtfully selected because they are believed to support the attainment of those goals. The intended alignment between inputs and desired outcomes is also clearly documented through a comprehensive theory of action (TOA). A TOA is a rational argument for a proposed system design. It outlines, at a minimum, the components of the system, how they are related, and the mechanism by which they are intended to provide for intended goals. A TOA for a state school accountability system should articulate not only the goals of the system and intended use of system results, but also the resources and inputs provided by the system and how they are related; the actions (by districts, schools and educators) the resources are intended to support; and the means by which intended actions are expected to provide for desired outcomes.

It is the integrity with which the elements underlying a state's theory of action are operationalized and implemented within the context of an accountability system and the extent to which those elements provide for intended and unintended outcomes that are the focus of system evaluation. In effect, the evaluation of an accountability system examines the assumptions the program makes about: a) the extent to and quality with which desired actions occur; b) the validity and relevance of the system's measures; c) the appropriate use and interpretation of results by stakeholders; and d) the extent to which desired actions produce desired results. Research reflecting the degree to which these assumptions hold provides evidence defending the value and efficacy of the system. In addition, a well-designed evaluation can support the identification of system flaws, facilitate modifications that improve the quality of system-based outcomes, and increase system efficiency.

Determining the focus of accountability system evaluation efforts can be a daunting task. One must determine the type of evidence that is needed to evaluate the quality and impact of the system, and also prioritize data collection efforts in light of limited time, money and staff. At a high-level, there are two types of evidence that can inform claims regarding the quality, fairness and utility of an accountability system: 1) descriptive evidence that demonstrates the coherence of the system, and 2) evidence of efficacy that documents the extent to which the system and its components are contributing to the attainment of goals in the manner expected.

Descriptive evidence is documentation that clearly outlines the goals of the accountability system and the mechanism by which the design of the system is anticipated to achieve those goals. It describes each component of the system design, the expected relationship between components, and the manner in which they jointly support the attainment of system goals (see Figure 1). It articulates key indicators of success, the measures by which they are operationalized, methods for aggregating data within and among indicators, and the process by which ratings are generated to report system-level performance. Furthermore, descriptive evidence should communicate the expected impact of the system on different stakeholders, the manner in which results from the system are intended to facilitate that impact, and the assumptions that must be met for the desired outcomes to be realized. In effect, descriptive evidence provides a detailed account of the system design, intended actions, and desired results in a manner that serves to highlight those aspects of the accountability system (e.g., assumptions, measures, processes) which should be the focus of evaluation (and, by omission, those not *necessary* for evaluation).

Figure 1. Descriptive Evidence



Evidence of efficacy supports claims about the extent to which and quality with which: 1) system components (e.g., inputs, resources, supports, consequences) are implemented as intended; 2) identified indicators and their associated measures provide for relevant, reliable and defensible information aligned to the outcomes of interest; and 3) system components are having the impacts intended. In effect, evidence of efficacy allows a program to evaluate the quality of implementation and the resulting effect on intended and unintended outcomes. This evidence may take many forms and necessitate a variety of data collection techniques (e.g., interviews, surveys, longitudinal analysis, and research studies). To illustrate, a sample of the types questions that may be addressed when collecting evidence of efficacy are provided below:

- Do identified indicators (i.e., growth, achievement, college readiness, etc.) provide valuable and consistent information to support decisions about schools?
- Are system-based measures (e.g., graduation rate, percent proficient, etc.) fair and reliable?
- Are the defined program inputs (e.g., professional development, curriculum materials, programs, interventions, etc.) and resources working in the manner intended to drive change?
- Are the measures selected for inclusion in the system relevant, accurate and useful given the indicators they are intended to represent? (For example, the use of high school graduation rate as an indicator of college readiness)
- Are districts and schools using system-based information as intended?
- Are specified consequences and supports having the impact expected? That is, are they having the desired effect on intended outcomes?
- Are negative consequences occurring and, if so, are they mitigated?

The Accountability System Program Evaluation Framework presented in Figure 2 outlines the information necessary to support the development of a comprehensive evaluation plan for a state accountability system. It includes: the System Goals, Theory of Action, Inputs/Resources, Intended and Unintended Outcomes, Program Measures, and Evidence Supporting Assumptions. The framework is intended to guide the evaluation design by a) establishing the inter-relationships within an accountability system that may be the focus of an evaluation; b) clarifying the intended and unintended inputs and outcomes that will be the focus of data collection; c) necessitating the articulation of key design assumptions; and d) guiding claims about program efficacy that may be made based on findings from the evaluation.

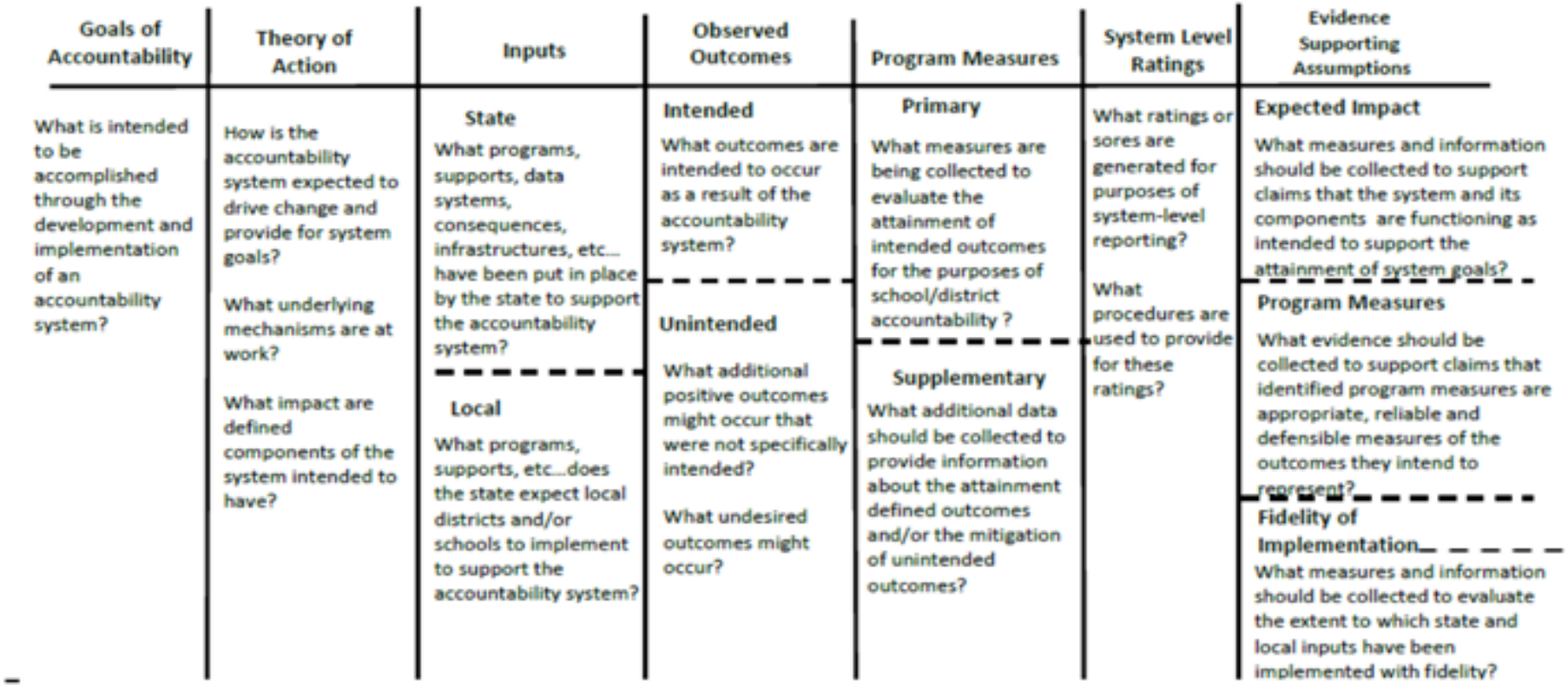
Each component of the framework is defined in terms of the key question it is intended to address and the horizontal presentation reflects the required coherence among components. This structure serves to highlight intended links among system components and, consequently, the assumptions underlying the assessment design. For example, the primary purpose of system evaluation is to determine whether an accountability system is doing what is was designed to do. For this reason clear specification of the Goals of the System is presented column 1. If the goals of the system are not transparent, the design of the system, as represented in columns 2-6, and the evidence necessary to evaluate its effectiveness (column 7) may not be appropriately specified.

The second column, Theory of Action, accounts for the fact that different states may have different hypotheses regarding the manner by which common system goals (e.g., all students leave school college and career ready) are achieved. Therefore, to evaluate whether a system is coherent and working as intended it is necessary to understand the state's hypothesis as to how the system is intended to bring about change. For example, to reach the goal that *all* students leave school college ready, a state's theory of action may be that the accountability system will:

- provide for data that better informs the specification of appropriate school and teacher supports and/or
- motivate teachers and schools to try harder than they have in the past by making school performance transparent and/or
- more appropriately focuses financial and human resources on those schools needing the greatest support and/or
- facilitate discussion and collaboration among teachers within and across schools in a way that improves teacher instruction and student learning.

Different theories regarding the role of the system in providing for CCR will lead to the selection and prioritization of different inputs, outcomes, and program measures. Furthermore, each theory is associated with a different set of assumptions about what motivates stakeholders and the activities/interactions necessary for the system to work as intended. For example, an assumption associated with the second bullet would be that educators are not currently putting forth their best effort and will improve as a result of imposed accountability. Since these types of assumptions drive the focus of evaluation (e.g., do teachers efforts improve when the system is put in place?) a state's values and beliefs related to role of the accountability system must be clearly understood.

Accountability System Program Evaluation Framework



DESIGN ASSUMPTIONS

- What assumptions underlie the proposed theory of action?
  - What conditions or actions are we assuming will drive intended change?
  - What factors are we assuming serve to motivate stakeholders as intended?
- What assumptions are being made regarding the role, impact and implementatio of inputs?
  - What actions are we assuming will be taken by whom (districts, schools, teachers)?
  - What conditions or necessary infrastructures are we assuming exist?
- What assumptions are we making about how Outcomes will be obtained and their relevance for making inferences about the attainment of prioritized goals?
  - What is the rationale for selecting the outcomes of interest?
- What assumptions are being made about the program measures selected to quantify the attainment of desired outcomes, with respect to: reliability, validity, accuracy, relationship to other measures in the system, availability, etc....
- What is the rationale for the procedures used to establish system level ratings? What assumptions underlie that rationale?

The third column, Inputs/ Resources, are the programs, policies, initiatives, interventions, data structures, etc. put in place at the state or local level to support the state's hypothesized theory of action. Some inputs will be specified and imposed by the state (e.g., auditing of schools identified as "at risk"), while others may be state-specified, but locally implemented (e.g., the development and maintenance of school improvement plans). For example, a theory of action suggesting that gains in readiness are attained through improved teaching practices that result from teacher collaboration (as implied by bullet 4) would be supported by providing teachers with frequent, high quality group-based professional development opportunities.

Outcomes (column 4) are the intended and unintended actions, behaviors, and consequences that result from system implementation. Outcomes are often expressed in terms of Indicators, which are subsequently operationalized in terms of Program Measures (column 5). For example, in most states the inclusion of an accountability system is intended to provide for gains in *student achievement* (i.e., the indicator) as measured by student *performance on state or nationally developed K-12 academic tests (i*.e., the Program Measure*)*. That is, student achievement is considered an important indicator of school performance that should be positively influenced, in terms of student performance on the state test, if the accountability system is working as intended.

Within the framework, program measures are defined as either Primary or Secondary. Primary measures are those associated with outcomes that are directly used to inform system-level ratings (i.e., included as part of the school/district accountability model). They typically include such things as test scores, participation rates, graduation rates, attendance and other measures considered to be reliable, comparable, and valid for making inferences regarding school performance. Secondary measures, on the other hand, are measures collected to monitor the broader impact of the system with respect to those outcomes for which schools/districts are *not* held directly accountable.

Program measures are often aggregated in a variety of ways to produce the ratings, scores or grades used to report system level results and identify schools in need of support. Since these results are the means by which the inputs and resources believed to be beneficial are assigned to schools, the procedures by which they are calculated and assigned, and the associated rationale, must be understood and evaluated, as reflected by column 6.

Finally, the last column, Evidence Supporting Assumptions, requires specification of the data/information necessary to lend credence to the assumptions underlying the system design. Within the framework evidence has been chunked into 3 categories - that related to assumptions underlying the:

- intended impact of the system (inputs, resources, results, etc…) on stakeholders, districts and schools;
- quality and appropriateness of program measures given the outcomes they are intended to inform; and
- fidelity of system implementation.

While many of the questions posed in this framework will have already been addressed during system design, addressing these questions with an eye to system *evaluation* requires a different perspective. In the former the focus is on selecting, defining and aggregating system measures in a manner that will support the attainment of system goals. In the latter the focus is on identifying the assumptions, conditions, beliefs and practices that must hold in order for the system to provide for valid inferences about the quality/performance of schools and the usefulness of the system and its components.

As a result, when thinking about the design components indicated in columns 2-6 is it useful not only to address the questions posed, but also ask:

- What is the rationale for this design decision?
- What conditions or assumptions must hold in order for this design feature (input, measure, etc…) to function as intended?

Specific questions targeted at highlighting the assumptions associated with each design component are provided at the bottom of the framework. The answers to these questions dictate, in large part, the evidence necessary to address the questions outlined in column 7.

For example, assume an aspect of a state's theory of action is that collaboration among educators is a key mechanism for achieving the goal of continuous school improvement. As a result, the state encourages districts to designate time for educators to meet and discuss key issues believed to be crucial to improving student/school performance (i.e., a local input). An assumption is that schools and districts not only provide this opportunity, but that educators participate and benefit from the experience in ways that positively influence teaching and learning. To support these assumptions, several types of evidence could be collected such as, documentation from schools/districts regarding the number of meetings convened, participation rates, and feedback from teachers regarding the quality and utility of the experience for improving instructional practices. In the absence of this type of information, it is impossible to establish that teachers collaborated and, consequently, make claims about the efficacy of collaboration for school improvement.

It is important to note that the framework is not intended to imply a one to one correspondence between different elements in the system. For example, there may be unintended inputs introduced at the local level or supported by third-parties (e.g., foundations or university-based research efforts) that have a positive, negative, or neutral effect on the intended outcomes. Similarly, there may be unintended outcomes that result from system implementation, but are not directly attributable to a particular input. In some cases, multiple inputs may influence a given intended outcome, and outcomes may be evaluated in terms of multiple program measures. When resources allow, stronger claims about the efficacy of an accountability system can be made when both intended and unintended inputs and outcomes are examined, and the combined effect of multiple inputs is considered.

Finally, the evaluation framework presents elements in a horizontal manner, but it is not implied that these elements are defined necessarily in a linear, sequential order during development. For example, intended outcomes (e.g., increased rates of growth, more students achieving college readiness) may be

defined prior to state/locally defined inputs. Similarly, program measures selected to operationalize a particular outcome (e.g., graduation rate) may be defined in advance of program implementation.

Applying the Framework

In this section, we illustrate an application of the framework through two closely related examples. . Taken as a whole, it describes elements of a potential evaluation approach in service to a goal commonly associated with school accountability systems: *increasing the percentage of students who graduate high school ready for college and careers*. The first example focuses more narrowly on evaluating the extent to which the accountability system has affected high school student college and career readiness. The second example expands the evaluation by focusing on the extent to which inputs have been implemented as intended and the effect they are having on outcomes.

We present these illustrations by outlining the TOA (relative to this goal), describing the primary design components, and articulating the evidence needed to evaluate the assumptions underlying the system design. However, in so doing we do not wish to imply that these activities are detached. In fact, a central theme of the framework is that design and evaluation are inextricably linked, which is why we must present the former to adequately illustrate a framework for the latter.

Example 1: Evaluating the Impact of the System on Desired Outcomes

Many states have prioritized college and career readiness as a central value for public education. To that end, policy makers are increasingly looking to school accountability systems to help promote and track attainment of this goal.

Naturally, the process must start with a clear definition of readiness and specific statements about the intended outcomes. For this example, one such statement may be: students will have the academic knowledge and skills determined to be important prerequisites for success in a specified introductory, credit-bearing college courses without remediation. We acknowledge this statement addresses a narrow portion of the full breadth of readiness. Indeed, much has been written about the construct of both college and career readiness which includes a wide range of skills, dispositions, and competencies a full exploration of which is beyond the scope of this paper (see, e.g. Conley, 2012). We digress here only to point out that for each readiness claim associated with the accountability system, a host of design and evaluation initiatives must be developed in support of that claim. For example, if the accountability system is also intended to promote and measure attainment of learning strategies associated with academic success, such as collaboration and persistence, additional definitions, measures, and evaluation activities would need to be developed to examine the extent to which the accountability system supports the attainment of this goal.

Once the goal is established, the next step is to specify the theory of action for how the system will promote that goal. A theory of action explicates the mechanism by which the accountability system will bring about the desired change. Developing this theory is an important step in designing the system and judging the degree to which the system is effective. For this first example, imagine a simple accountability system in which the state provides:

- timely reporting of results of the state tests, including total test scores and sub-scores, to districts and schools so that they can:
- examine results to inform modifications to their curriculum and areas in which instruction is in need of strengthening

In addition to the larger goal of improving readiness of all students, the system's theory of action might also establish conditions believed to support the achievement of the goal of readiness. Such interim goals may include:

- High schools offer more courses that more closely align with the academic expectations in college.
- More students enroll and successfully complete courses that will prepare them for academic success in college.
- Effective remediation/ support will be implemented for students who are not initially successful meeting performance expectations associated with academic success in college.

As depicted in Figure 3, the theory of action stipulates that test results will be used by schools and districts to inform modifications to their program of study in ways that better prepare students for college, and result in gains in readiness. Further, the theory of action stipulates that modifications to the program of study will occur in two ways: a) improving the rigor of courses; and b) providing additional support to low achieving students.

Figure 3: Theory of Action

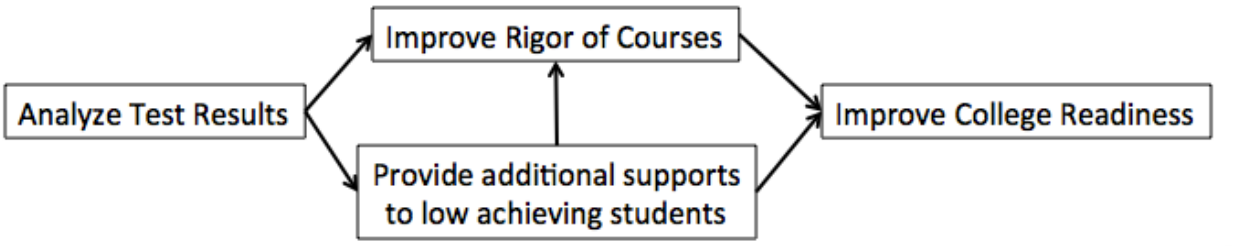
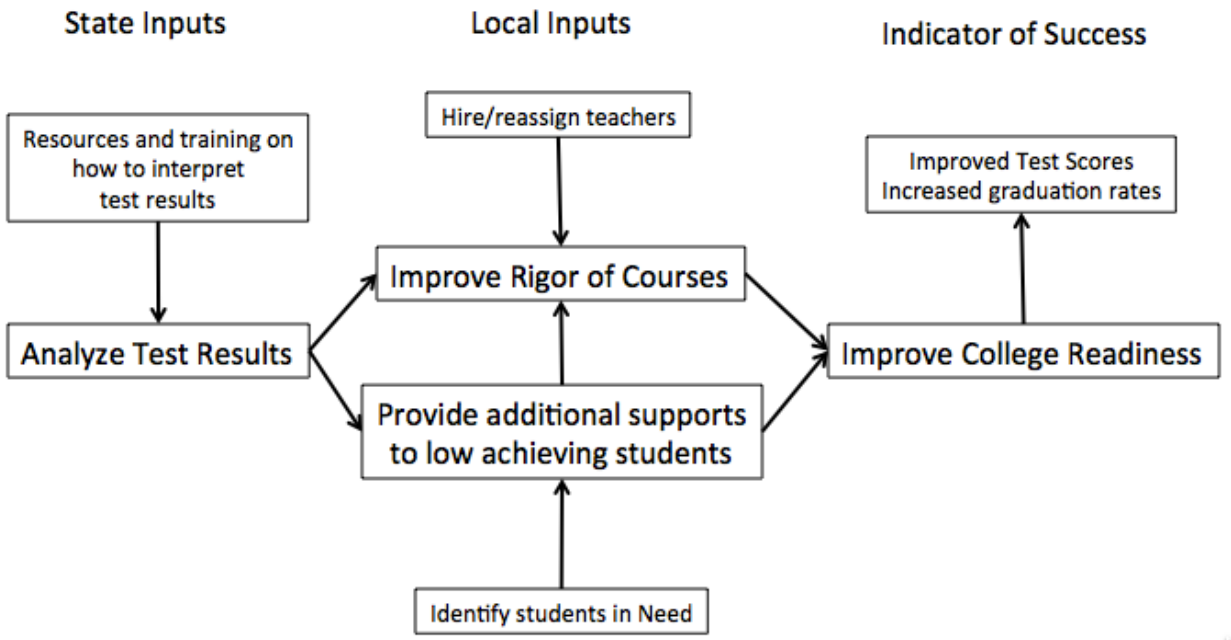


Figure 4 expands the depiction of the theory of action to show how the *state inputs* support each component of the theory. State inputs to the accountability system include provision of the resources and training necessary to support districts in using test results to inform the selection and implementation of new course options (i.e., college-prep and remedial). Local inputs include the addition of teachers and courses that serve to improve student access to college-prep course work and/or support students who require remediation.

Figure 4: State Inputs Supporting the Implementation of the Theory of Action



Because the education accountability system focuses on K-12 schools, outcomes indicating gains in college readiness and student achievement for elementary schools are measured in terms of student performance on the state test. At the high-school level, college readiness is evaluated in terms of these measures, in addition to graduation rate and the percentage of students achieving ACT's College Readiness benchmarks. For the purposes of this example, it is assumed that sufficient evidence was previously collected to establish that these measures are sensitive to instruction and provide for a valid indicator of readiness as defined above. If validity evidence that supports the use of these tests for this purpose had not been previously collected, the evaluation activities outlined below would need to be expanded to include the collection of such evidence.

Once the theory of action and design decisions in support of that theory are established, a plan for collecting evaluation evidence can be developed. This starts by clearly identifying intended and potential negative unintended outcomes. Then, the measures that will provide information about these outcomes are identified. For example, as noted in the previous section, the intended outcomes in this scenario are gains in college readiness and student achievement. Additional, secondary outcomes include: a) increase in the number of high school courses that address academic expectations for college; b) increase enrollment in these courses; and c) remediation and support for low-performing students. Sources of evidence for each of the final and interim outcomes should be identified. As an example sample measures for the final and interim outcomes listed in the example are provided in columns 4 and 5 of Table 1, which reflect key elements of the evaluation framework for this example.

Elements of Accountability Framework				
Goal of System	**Theory of Action**	**Inputs/ Resources**	**Outcomes (Indicators)**	**Program Measures**
Increasing the percentage of students who graduate high school ready for college and careers	Accountability system provides for test results that *inform* the establishment of new, more rigorous, courses; and/or facilitate the provision of remedial support to those at risk.	**State Inputs:** Professional development that supports districts/schools in the use of assessment results to inform decisions about new course/program offerings. Financial resources to support the addition of new courses/ remediation **Local Inputs:** Additional courses aligned to college-level expectations. Remediation courses/opportunities for students at risk of not being college ready.	**Intended**: Increased levels of college readiness Improved student achievement **Intended Secondary Outcomes:** Increased course/ program offerings - rigorous courses aligned to college-expectations - support programs that provide for remediation Gains in course participation **Unintended /Negative Outcomes** Fewer students enroll in courses associated with college readiness More students drop-out of high school Overemphasis on testing narrows the curriculum	**Primary** (Used in system) - Percentage of students achieving college-ready, or on track to college-ready, standards on state test. - Percentage of student achieving ACT CR benchmarks - Graduate rate* - Gains in average performance on state test. **Secondary** - Number of college prep and AP courses - Audit of curriculum/ syllabi for college preparatory courses - The number and type of support programs offered by course - Number/percent of students enrolled in new rigorous courses - Number/percent of students not meeting readiness standards who participate in support Programs - Annual trend data reflecting enrollment of students in college-prep courses - Drop-out rates by district, school, and subgroup - Audit of curricular materials (e.g. syllabi) in courses - Focus groups and surveys of educators - Trend data for measures in not included in the accountability system

In addition to evaluating the intended outcomes, it is also important to identify unintended, negative consequences and collect evidence to assess these threats. Examples of these outcomes and their associated measures have also been provided in Figure 4.

Evaluating Relationships between Interim and Final Outcomes

The measures listed above can be used to indicate the extent to which intended and unintended outcomes are occurring. For example, an increase in the average test scores for the majority of schools provides evidence (i.e., indicates) that student readiness is improving. Similarly, increases in the number of courses addressing college expectations provides evidence that schools are responding to the accountability system by offering students more courses that are designed to improve student readiness. However, these two data points alone are not sufficient to conclude that the theory of action is functioning as designed. To support this claim, one must collect evidence supporting the assumption that program-based inputs (i.e., college prep courses) are directly related to the outcomes of interest (increased college readiness and achievement).

For this example, the assumption being made is that the inclusion of college preparatory courses will provide for improved student performance. One approach to examining this assumption may be to compare changes in observed outcomes between schools that differ with respect to the attainment of these interim goals. As an example, one might classify schools as those that have increased the number of courses offered that address college expectations and those that have not. Mean changes in test scores might then be compared between these two groups of schools. Cases where the mean change in test scores is larger for schools that have increased courses addressing college expectations than for schools that have not increased these courses, evaluation evidence might support the claim that there is a relationship between increasing the number of these courses and improving student readiness. Of course, more sophisticated approaches could be used to examine other factors that may be impacting this relationship, such as the number of such courses introduced, the percentage of students taking these courses, the quality of instruction and curriculum, the presence of remedial programs, and other socio-economic characteristics of the schools. The larger point is the importance of examining assumed relationships among desired outcomes and the design components intended to provide for their attainment. This is necessary to develop an understanding of whether the factors believed to affect improvements in readiness, as presented in the theory of action, are having the intended effects.

To extend this point even further, Table 2 outlines some of the key assumptions underlying this simple design for which evidence should be collected to determine if the TOA is functioning as intended.

Assumptions			
Theory of Action	**Inputs/ Resources**	**Outcomes**	**Program Measures**
Data is provided in a timely manner that allows for the use of results as intended.	Professional development is offered in a manner that supports broad participation.	Identified outcomes/ indicators provide for valid inferences regarding student	Identified measures are related to or represent the indicator of interest Measures are related to

Districts, schools and educators use the accountability system results to inform instructional decisions. (i.e., identify students requiring remediation; inform course offerings, etc…). Students take advantage of new course offerings and/or remediation opportunities.	Stakeholders participate in PD opportunities. PD is considered useful and of high quality. Rigorous courses, when taken, improve student performance and readiness Support programs are effective and serve to improve readiness	readiness upon exit from HS (as defined within the context of this state system)	each other in a manner consistent with that which would be expected. (e.g., attainment of ACT benchmark and state CCR standard). Measures are reliable at a level consistent with their intended use.

Beyond examining the intended and unintended outcomes and the relationships among design components and final goals, it is also important to consider the extent to which system-level results are reliable and provide for valid inferences related to the attainment of prioritized goals. Clearly, one factor that may affect the validity and reliability of the system- level results is the quality of the program measures; a test with low reliability is likely to result in unstable patterns of change over time. However, there are other factors that should be considered when examining the validity and reliability of system-based outcomes. Approaches to examining reliability and validity of program outcomes are examined in greater detail in a subsequent section.

Summary of Example 1

Through this simple example, we see the importance of defining the program goals, theory of action, inputs, and outcomes, and the assumptions underlying their interactions to guide the design of an evaluation. We also see the separation between the outcomes and the measures of the outcomes. And we explored how unanticipated outcomes can be identified and examined. Finally, the value of examining relationships among design elements and final goals is demonstrated. Collectively, each component of the evaluation contributes to a holistic judgment about the extent to which the program is achieving its goals and allows a program to balance those intended effects with unintended outcomes.

Example 2: Relationship between Inputs and Outputs

This example builds on the first example by expanding the theory of action and the inputs provided through the accountability system. This example then describes how an evaluation may examine the extent to which inputs are put into practice in a quality manner and the resulting relationship between program inputs and outcomes.

Design

In this example, the accountability system uses the same summative tests as the first example to provide a measure of college readiness. The program also provides the same timely reports that it expects districts and schools to examine in order to improve curriculum and instruction. In addition to the first example, the program's theory of action also holds that access to assessment content and professional development will deepen schools' understanding of what students are expected to know and do and at what level students are expected to demonstrate their abilities. Access to these resources are also expected to be used by schools and districts to modify their practices to help students meet these expectations and to monitor progress towards these expectations. To these ends, the state provides access to:

- Released test content including items, tasks, scoring rubrics, and exemplars for open-response items
- Interim assessments and an item bank that educators can use to build custom interim or formative assessment instruments
- Professional development resources and supplemental professional development funds.

In this system, these additional resources are provided to districts and schools so that they can:

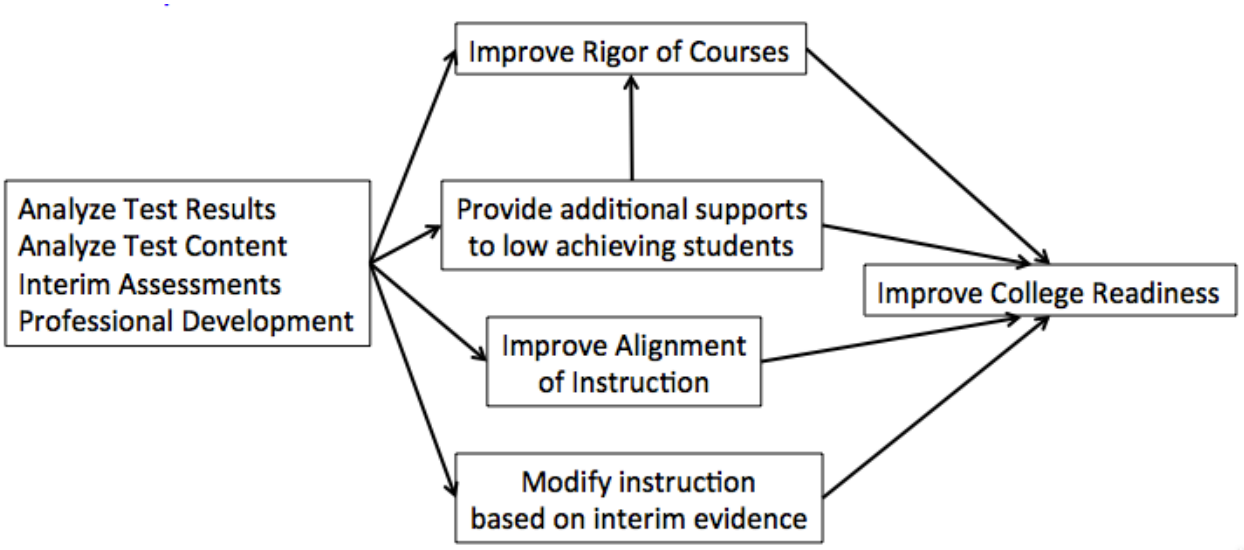
- Examine released content to deepen their understanding of what students are expected to know and do, and to then use this content to inform the content addressed during instruction
- Assess student achievement as knowledge and abilities are developed and to monitor progress towards readiness periodically throughout the school year
- Deepen educators understanding of content and performance expectations and modify curriculum and instructional practices to support student achievement of these expectations.

Additional interim goals established by this program include:

- More instructional time focused on developing the knowledge and abilities associated with readiness
- More frequent modification of instruction based on formative and interim evidence of student learning

As depicted in Figure 5, this more complex accountability system is based on a theory of action that stipulates that test results, test content, and professional development resources will be used by schools and districts to modify their program of study in order to better prepare students for college which will result in improved test performance overtime. Further, the theory of action stipulates that modifications to the program of study will occur by: a) improving the rigor of courses; b) providing additional support to low achieving students; c) improving the alignment of instruction with content and performance expectations; and d) using assessment data throughout the school year to identify and address learning needs of individual or sub-groups of students.

Figure 5: Theory of Action for Example 2



Evidence

As noted in the first example, the final intended outcome is an increase in scores on tests that provide measures of readiness. Interim intended outcomes include: a) increased number of high school courses that address academic expectations for college; b) increased enrollment in these courses; c) remediation and support for low-performing students; d) improved alignment of instruction; and e) increased use of formative and interim assessment information to tailor and/or individualize instruction. Sources of evidence for the final and first three interim outcomes are the same as the first example. Outcomes and measures for the additional interim outcomes may include those listed in Table 3:

Outcome	Measure(s)
Improvement in alignment of instruction	- Survey of educators focusing on time spent addressing key content addressed in the standards - Audit of a samples of program curricula by qualified experts - Case study of instructional practices in a sample of schools
Increased use of formative and interim assessment information to tailor instruction	- Survey of educators on frequency of use of formative and interim assessment instruments - Review of downloads and/or use of assessment content resources on state web-site - Review of assessment administrations on state-provided formative/interim assessment system
Use of professional development resources and	- Review of requests by and allocation of funds

funds to improve educator understanding of expectations and improve instructional practices	to schools/districts for supplemental professional development - Survey of schools/districts about how supplemental PD funds were used - Survey of educators about quality of PD provided through supplemental funds - Survey of educators about how they modified instructional practices based on PD

The theory of action in this example is more complex than the first in two important ways. First, the state provides a larger number of resources that are intended to impact programs and practices. In this example the state not only provides timely reports but also released content, interim assessments, and funds for professional development. Second, districts and schools are provided flexibility in how they use these state-provided resources. As an example, one district might use the state provided interim assessments to put in place a regimented interim assessment program that all teachers are required to follow, while another district may give educators the choice of when and how the interim assessments are used. Similarly, one district may opt to use professional development funds to support all-day workshops during which external experts come into schools to provide consultation. Another district may use the funds to provide release time for educators to work together to modify their instructional programs. And a third district may opt not to use the funds at all. Given the increased complexity and resulting flexibility inherent in this accountability system, an evaluation of the system might focus on the ways in which inputs are utilized and the relationship these uses have with outcomes.

As one example, the theory of action posits that the use of formative and interim assessment tools provided by the state will impact the extent to which educators tailor instruction to meet the needs of individual and sub-groups of students. The theory of action further stipulates that tailoring of instruction will have a positive effect on student readiness as measured by test scores. To examine this complex relationship, an evaluation might survey educators about their use of the state-supplied formative and interim assessment tools and use data records from the system to identify educators who use these resources frequently and those who do not use them at all. Recognizing that other formative assessment tools may be used to inform tailored instruction, the survey might also ask about the use of non-state-provided formative and interim assessment tools, and further divide the non-users of the state supplied tools into those who frequently use other tools and those who do not. Given that the intended effect of the use of these tools is the tailoring of instruction, the survey might also collect information about how educators use assessment results to identify those who tailor instruction from those who do not. In effect, this categorization results in six potential groups of educators as shown in Table 4: .

Table 4: Formative and Interim Assessment Educator Use Categories

	Frequently Tailor Instruction	Infrequently Tailor Instruction
Frequent Use of State Tools	Group 1	Group 2
Frequent Use of Non-State Tools	Group 3	Group 4
Infrequent Use of Tools	Group 5	Group 6

This categorization can be used to examine the extent to which there is a relationship between the use of formative and interim assessment tools and the tailoring of instruction, and whether this effect differs between the use of state versus external assessment tools. In effect, this analysis allows the evaluation to examine the extent to which the theory that use of assessment tools influences tailoring of instruction holds. If it does, additional analyses might examine the extent to which the use of assessment tools and subsequent tailoring of instruction is related with the outcome measure – namely test scores. This analysis allows the evaluation to examine the extent to which tailoring of instruction based on formative and/or interim assessment improves student readiness as measured by the state tests. If this component of the theory holds, the evaluation might examine whether the effect of using the state-provided tools differs from the use of non-state assessment tools. Through this analysis, the evaluation may help the state determine whether the provision of the formative and interim assessment tools is bringing value above that which schools and districts can achieve independently of state support. Of course, this analysis might also consider whether districts and schools that make use of the state-provided assessment tools would be positioned to access the non-state tools if the state tools were not available; an analysis that might help inform the continuation of the state-provided tools regardless of their comparative effect.

As this example demonstrates, as the number of links in the chain of logic specified in the theory of action increases, the complexity of the evaluation design also increases. This complexity, however, is important to tease out the extent to which state-provided inputs are being used by schools and districts and whether this use is having the intended immediate and long-term effects. This information can then be used to inform modification to specific components of the accountability system or, in some cases, elimination of a given component.

Summary of Example 2

Through this more complex example, we see the importance of identifying the immediate and long-term effects of each component of the accountability system. We also see the value in examining the extent to which each link in the theory of action holds and whether the chain of logic expressed by that theory holds. Finally, we explored the importance of considering the effect that external inputs might have on outcomes and comparing effects between system-provided and externally acquired inputs. This more complex evaluation provides a more robust understanding of the effects that each component of the system has on the intended outcome and places these effects in the context of other practices and resources employed by districts and schools. Collectively, this more complex and robust view better positions decision-makers to modify and improve their accountability system.

As in the first example, it is important to identify unintended, negative consequences and to collect evidence to assess these threats. For this example, potential threats and the associated measures would likely be the same.

Reliability and Validity of Program Effects

This section focuses on the reliability and validity of program effects. It is important to note that the focus is on program effects rather than on the tests used as outcome measures. As indicated above, the validity and reliability of tests used as outcome measures is an important factor that may impact the reliability and validity of program effects. For the purposes of this discussion, it is assumed that prior analyses have established the reliability and validity of using the state's tests as a measure of readiness; thus this section will not address approaches to examining test reliability and validity.

When considering the reliability of program effects, there are two important questions to address: 1. Are effects consistent over time; and 2. Are effects consistent within and between schools and student sub-groups. Past research has found larger fluctuations in changes to outcome measures across years for small schools as compared to large schools. The first question examines the extent to which instability in outcomes across years threatens the value of classifications made by the accountability system based on outcome measures. In cases where large fluctuations in classifications are frequent – one year a school is deemed high performing and the next low performing or vice versa – the reliability of such classifications is threatened. In turn, low reliability of classifications may negatively impact the believability or trust in the accountability system by schools and the public. Where fluctuations occur, it is important to explore factors that may be related to these fluctuations, such as school size or populations served by school. When associated factors are identified, the evaluation findings may be used to modify the accountability system to minimize the influence of these external factors. As an example, if a correlation between school size and instability of classifications based on one-year results is found, the system might adjust the approach used to classify schools by employing a multi-year average performance rather than single-year performance.

Similar analyses might also example the extent to which results are consistent among schools with similar characteristics. 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. As an example, one would expect that schools that serve a similar percentage of students who are developing English language proficiency would require similar levels of remediation for English language arts. In turn, one would expect that schools that form this group would establish remedial programs at similar rates and the number of students served by these programs would be similar. Dramatic shifts in outcome measures for schools or differences among schools that form a given group will signal a lack of stability that will erode the credibility of the outcomes.

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?

The first question addresses the extent to which the model differentiates outcomes among schools. 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 relationship between school type (e.g. high schools, middle schools, elementary schools) or schools size and accountability outcomes, this works against the credibility of the model as these factors should not be strongly related to school quality. This is not to say that overall (e.g. mean) differences across different types or sizes of schools signal a validity threat. Rather, the distribution of outcomes should generally span the full range for factors not directly tied to performance. For example, there should be a good spread of performance that spans the range of outcomes for small, medium, and large sized schools, even if small schools generally outperform larger schools.

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 schools where a higher percentage of teachers who are National Board certified to receive favorable outcomes. Similarly, high schools with higher graduation rates or higher college-going rates should, in general, receive more favorable outcomes than 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.

Appendix

The Wisconsin Accountability System: A Case Study

To provide additional clarity and context for applying the evaluation framework, we present a case study based on Wisconsin's current school accountability system. We start with a review of the process that established the foundation for the system and clarified the goals and theory of action. Then, we focus on the intended outcomes and the program measures selected to track these outcomes. Finally, we illustrate some potential sources of evidence as part of an ongoing monitoring and evaluation process. Throughout the case study narrative, we highlight linkages to the evaluation framework in shaded boxes.

Background: Determining Goals and Theory of Action

The design of Wisconsin's current accountability largely reflects a shift in accountability priorities resulting from discussions in the state in 2011. That year, heeding calls for a Wisconsin-specific accountability system, the State Superintendent of Public Instruction, Governor, and chairs of the Senate and Assembly education committees convened a School and District Accountability Design Team. This group, comprised approximately 30 education stakeholders representing various education entities, school and district roles, and student populations, discussed key goals and principles of an accountability system "of and for" Wisconsin.

According to the Accountability Design Team, a quality accountability system will:

- Support high-quality instruction in all publicly funded schools and districts;
- Include all publicly funded students in accountability calculations;
- Measure progress using both growth and attainment calculations;
- Make every effort to align this work with other state educational reform initiatives;
- Align performance objectives to career and college readiness;
- Focus on and include multiple measures of student outcomes that can be used to guide and inform practice and for accountability purposes;
- Use disaggregated student data for determinations and presorting to facilitate the narrowing of persistent achievement gaps;
- Make valid and reliable school and district accountability determinations annually;
- Produce reports that are transparent, timely, useful, and understandable by students, parents, teachers, administrators, and the general public;
- Provide differentiated systems of support to the lowest performing schools and districts including professional development targeted to their deficits;
- Recognize the highest performing schools and districts, and disseminate their best practices to schools serving similar populations to help scale up high performance statewide;
- Have reasonable and realistic implementation goals that ensure the state, districts, and schools have the capacity to fully implement the accountability system and act on the results; and

Priority Goals

The goals for Wisconsin's school accountability system are evident in the focal areas determined by the Accountability Design Team. The system is designed to promote: student achievement, academic growth, equity of outcomes, and readiness for post-secondary success.

- Remain open to feedback and findings about potential system improvements through implementation to ensure maximum effectiveness of the system.

Ultimately, the Design Team identified the four key areas of focus for the accountability system:

1. Student achievement
2. Student growth
3. Closing gaps
4. On-track to graduation and postsecondary readiness

These came to be known as the report card's Priority Areas and reflect the systems' goals.

The group also felt that the accountability system should engage multiple measures that reflect a value placed on varied postsecondary outcomes. They wanted the system to focus not only on English language arts and mathematics assessment performance, but also science and social students and 21st century skills as appropriate data become available. It also stated that college and career readiness should be measured differently for elementary and middle schools than high schools.

The principles and recommendations laid forth by the Accountability Design team provided an initial framework for more detailed design of the accountability measures and reports. The Design Team discussions also informed the high level Theory of Action (TOA) for how the system was intended to promote the identified goals. The TOA posits that designing and producing school and district report cards that treat every school as fairly as possible, are valid, reliable, and transparent, will inform local improvement planning and highlight actionable areas of performance that reflect key values in the educational system. Moreover, appropriate supports and interventions that are based upon a continuum of levels of support, directly linked and adjusted according to accountability ratings, will help support the intended goals.

High Level Theory of Action

Report cards inform local improvement planning and highlight actionable areas of performance that reflect key values in the education system. This influences a continuum of support initiatives linked to accountability outcomes.

Measures and Design Features[1]

The school and district report cards include the four priority areas identified by the Accountability Design Team, as well as three Student Engagement Indicators, which reflect individual measures of importance that, to some extent, reflect on the validity of the priority area measures. Supplemental data play a key part in the report cards, in an effort to encourage those viewing the report card to "drill in," ask further questions, and ultimately attend to other, related data sources not captured in the report cards, such as local data. For members of the public that view the report cards, the data therein are intended to provide an understanding of overall performance in key areas.

[1] For additional information about Wisconsin's accountability system see:

- Report Card Technical Guide: http://dpi.wi.gov/sites/default/files/imce/accountability/pdf/School%20Report%20Card%20Technical%20Guide%202014.pdf
- Report Card Interpretive Guide: http://dpi.wi.gov/sites/default/files/imce/accountability/pdf/Interpretive%20Guide%202014.pdf
- Additional Resources: http://dpi.wi.gov/accountability/resources

Priority Areas

The priority areas were listed in the previous section and serve to clarify the intended outcomes. To track these outcomes, the following program measures are produced.

Student Achievement

Purpose: to show how the students' level of knowledge and skills at a specific district or school compares against state academic standards.

Measure(s): a composite English language arts (ELA) and mathematics performance of all students. The score is based on how students are distributed across the four WSAS performance levels, and it takes three years worth of test data into account.

Supplemental data: performance by subgroup.

Details:

- The method for calculating each content area score is based on assigning points to each of the district or school's students in each of the three measured years according to the student's performance level in that year. A student is assigned no points for being at the Minimal Performance level, one-half point for being at the Basic level, one full point for Proficient, and one-and-a-half points for Advanced.
- ELA and math are equally weighted, comprising 50-points each of the 100-point priority area score.
- For each year, students' scores are pooled to produce a district or school average. From those yearly averages, a three-year average is calculated. The averaging processes used in the calculations give greater weight to more recent years' data and also reduce the effect of year-to-year enrollment variability on aggregated test data. The score for each content area reflects this three-year average.

Program Measures

To track the prioritized outcomes program measures for Wisconsin's system include:

- Weighted index of ELA and mathematics performance on state tests
- Academic growth based on achieving target Student Growth Percentile (SGP) values
- Gap closure for identified groups based on improvement in test scores and/or graduation rate that exceeds comparison group
- Graduation rate
- Attendance rate (selected schools)
- Other academic measures associated with readiness or on-track to readiness
- Test participation
- Drop-out rates

Student Growth

Purpose: to give schools and districts a single measure that summarizes how rapidly their students are gaining knowledge and skills from year to year. In contrast to Student Achievement, which is based on the levels of performance students have attained, Student Growth focuses on the *pace of improvement* in students' performance. Student Growth rewards schools and districts for helping students reach higher performance levels, regardless of a student's starting point.

Measure(s): the heart of this measure is a point system that rewards schools and districts for students' progress toward higher performance levels from wherever they started. The point system also penalizes for student performance that regresses below the proficient level. The measure also rewards schools and districts that are already doing well by maintaining the high performance of their students, thus recognizing that

very high performing students may not be able to grow as much or as quickly as other students as demonstrated by results on the state assessment.

Supplemental data: growth by subgroup

Details:

- Unlike Student Achievement, the Student Growth Priority Area only reflects the progress of students taking the general education assessment because the scoring scale of the alternate has not permitted growth calculations.
- This score reflects the degree to which students are on target to move from their starting scale scores to higher (or lower) performance levels within a three year period, based on their Student Growth Percentile (SGP). Students' starting scale scores are taken from the year prior to the current year of test results and an individual SGP is calculated for each student. Points are assigned to students based on a comparison of their SGPs with target SGPs for higher or lower performance levels.
- Target SGPs represent the pace of growth a student would have to exhibit to be considered on target to reach a different performance level within the three-year measurement period. Usually, this reflects growth to a higher level within three years or decline below Proficient within one year. Target SGPs are calculated using data about the growth track records of preceding groups of students who shared a similar achievement history with the student in question.
- Separate scores are calculated for ELA and mathematics and then combined.

Closing Gaps

Purpose: The purpose of this Priority Area is to provide a measure in sync with the statewide goal of having all students improve while closing the achievement gaps that separate different groups of Wisconsin students. It reflects the fact that achievement and graduation gaps are a statewide problem, not something limited to a small number of individual schools. The Closing Gaps Priority Area is designed to reward schools and districts that help close these statewide achievement gaps.

Measures: For this Priority Area, target racial/ethnic groups (Black students, Hispanic students, Asian/Pacific Islander students, and American Indian students) within a district or school are compared to White students statewide, their complementary comparison group. Students with disabilities, English language learners, and low-income students within a district or school are also compared to their complementary, statewide comparison group. A composite group (aka 'supergroup') is formed to meet the group size requirement (N=20) by combining at least two of the three above target groups when they do not meet the size requirement on their own. The Report Cards give credit for raising test scores and graduation rates for target groups faster than their statewide comparison groups. As a result, this measure encourages performance that lifts the performance of traditionally lagging groups, contributing to closing the statewide performance gaps.

Details:

- There are two components in the Closing Gaps priority area: Achievement Gaps and Graduation Gaps. If both apply for the district or school, each component score counts for half of this Priority Area score. If only one applies, the score for that component is the score for this Priority Area.
- The calculations for each of the two components follow the same basic procedure: Change in performance over the most recent three to five years is measured for each target group in the district or school and compared to the change in performance of the statewide comparison group. Change in performance is determined by finding the overall trend in performance, while also taking into account yearly fluctuations in enrollment. A minimum of three years of

performance data are considered, and up to five years are included when available. The difference between the group change and the statewide change is then calculated, producing the closing gaps indicator for each target group. The indicators from all target groups are then combined to produce an overall Closing Gaps score for that component.

- For the Closing Achievement Gaps component, performance means achievement in reading and mathematics, measured in the same way as for the Student Achievement Priority Area, except that students are pooled by group and not the entire district or school.
- For the Closing Graduation Gaps component, performance is measured with the four-year cohort graduation rate. Because Wisconsin began reporting cohort graduation rates in 2009-10, graduation data prior to 2009-10 are not available.

On-Track to Graduation and Postsecondary Readiness

Purpose: The purpose of this Priority Area is to give schools and districts an indication of how successfully students are achieving educational milestones that predict postsecondary readiness.

Measures: This Priority Area has two components. The first component is either a graduation rate—for schools that graduate students (i.e. high schools)—or an attendance rate for schools with no 12th grade. For most districts, both attendance and graduation scores will be included. The second component is a set of measures that include third grade reading achievement, eighth grade mathematics achievement, and ACT participation and performance, as applicable to the school. The scores for these two components are added to produce the Priority Area score.

Supplemental Data: subgroup performance

Details:

- Calculations for this Priority Area are based on an "all students" group.
- Component 1: Graduation Rate or Attendance Rate.
    - For schools that graduate students, a graduation rate is used as the indicator. For other schools, an attendance rate is used. Districts use both the graduation rate and attendance rate. Graduation rates and Attendance rates are highly correlated and have virtually identical distributions.
    - The graduation rate is the average of the four-year and six-year cohort graduation rates.
    - The attendance rate is the number of days of student attendance divided by the total possible number of days of attendance. The attendance rates of the "all students" group and the student group with the lowest attendance rate are averaged to produce the report card attendance rate.
    - The performance on this component accounts for a fixed 20 percent

- Component 2: Other On-Track Measures.
    - A school and district may have up to three 'Other On-Track' measures contributing to the score for this component: a third grade reading achievement indicator, an eighth grade mathematics achievement indicator, and a combined ACT participation and ACT performance indicator.
    - Third grade reading achievement and eighth grade mathematics achievement are measured in the same way as in the Student Achievement Priority Area.
    - The ACT Participation and Performance score is the average of five rates for twelfth-graders: the ACT participation rate and the college readiness rates for all four ACT subject areas.

- A composite score for this component accounts for a fixed five percent of the weighted average priority areas score, regardless of, overall, how many Priority Areas apply to the school.

Student Engagement Indicators

Three performance indicators measuring student engagement are vital indications of school and district effectiveness. Low test participation reduces the validity of any comparisons and conclusions that can be drawn from assessment data. High absenteeism and dropout rates point to other educational shortcomings. Because of the significance of these three indicators, districts and schools that fail to meet statewide goals marking acceptable performance will receive fixed deductions from the weighted average priority areas score.

Approaches to Ongoing Monitoring and Evaluation

We conclude this case study with some suggestions for potential evidence that may be collected to evaluate selected elements of the Wisconsin school accountability system. The sources of evidence shown in the following table are not intended to be comprehensive. Rather, this is intended to illustrate elements of the evaluation framework.

Component	**Potential Sources of Evidence**
Expected Impact	- Trends in student performance on state tests overall and by subgroup - Annual changes in magnitude of achievement gaps for academic measures and graduation rate - Percent of students enrolling in credit-bearing college courses - Increased student engagement as measured by attendance and absenteeism - Use of data to inform local decisions increases - Local decisions related to behavioral supports, curriculum, or staffing (for example) are adjusted based on, in part, performance as measured by the accountability system -
Program Measures	- Indicators are stable (e.g. year-to-year growth outcomes are positively correlated) - Outcomes are not correlated with unrelated factors (e.g. correlation between growth and prior-year status is low)
Fidelity of Implementation	- Focus groups reveal that reports are clear and helpful Surveys show that educators use results in planning and improvement efforts

We stress that ultimately the value of an accountability system is tied to the extent to which it both incentivizes the desired behaviors and produces information that stakeholders can and do use to improve student achievement. In the best case, these claims are made clear in the theory of action and are put to the test in the evaluation process. For example, if the theory of action holds that high school educators will provide instruction on more challenging academic content to prepare students for college, evidence to support this claim might include: review of syllabi or focus groups with teachers. As another example, if the theory of action in Wisconsin holds that support strategies triggered by the

system, such as providing supplemental educational services, will be effective, a study designed to compare similarly performing students who do and do not receive the services will help the state determine if these strategies are producing the desired result.

Ideally, evidence is collected, evaluated, and documented each year and the model will be refined as needed. In this manner, states improve the likelihood that the accountability system works to support the intended goals.