November 2017

State Systems of Identification and Support under ESSA:

Evaluating Identification Methods and Results in an Accountability System

CCSSO
Council of Chief State School Officers

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

State Systems of Identification and Support under ESSA:
Evaluating Identification Methods and Results in an Accountability System

Juan D'Brot, Ph.D.
Susan Lyons, Ph.D.
Erika Landl, Ph.D.

One Massachusetts Avenue, NW, Suite 700 • Washington, DC 20001-1431
Phone (202) 336-7000 • Fax (202) 408-8072 • www.ccsso.org

CONTENTS

INTRODUCTION

This paper is the second in a series of two discussing systems of school identification and support under the Every Student Succeeds Act (ESSA). This paper focuses on both technical and policy considerations for states in evaluating the success of their identification and accountability systems under ESSA. In an effort to inform readers on evaluating accountability systems, it describes considerations and methods to (1) evaluate the reliability of accountability scores and performance designations—including those schools identified for support and improvement, (2) evaluate the utility and impact of accountability systems in general and school identification systems in particular, and (3) evaluate the link between local behaviors and outcome improvement driven by accountability. This paper is intended to supplement the first in this series, *State Systems of Identification and Support under ESSA: A focus on designing and revising systems of school identification* (Lyons, D'Brot, & Landl, 2017). The first paper provides a comprehensive overview of the federal law and key design considerations for states as they develop and revise their systems of school identification under ESSA.

INTRODUCING A THREE-STEP APPROACH TO EVALUATING ACCOUNTABILITY SYSTEMS

This paper recommends an ongoing evaluation of states' accountability and support systems and offers concrete strategies to do so. We recommend that this process include examining (1) the reliability of accountability scores and designations, (2) the utility and impact of accountability systems, and (3) the link between behaviors and outcome improvement driven by school identification and accountability. These three approaches are briefly introduced before being described in detail in the remainder of the paper.

Reliability in Accountability Systems

Evaluating the reliability of accountability scores and school designations begins with understanding the impact of measurement and sampling issues on school-level estimates of system indicators. This can have an impact on the variability of the indicators and overall accountability system decisions. Once decisions are made, it is important to examine historical data or model data into the future to determine the consistency of school classifications. Finally, states can evaluate reliability by confirming that indicators interact appropriately and how schools are rated and grouped over time.

Evaluating Accountability System Decisions and Intended Outcomes

When making claims about the utility and impact of accountability systems, a clear understanding of system dependencies is necessary. That is, what kind of intended (or unintended) triggers or consequences in the identification process lead to other decisions in the system? Understanding how early decisions affect school designations can help identify unintended negative decision points in the accountability system.

We also recommend that state education agencies (SEAs) attend to how exit criteria are defined, how they relate to observed changes over time, and how leading indicators or upstream data are linked to long-term information or downstream data typically used as outcomes. This can support a stronger understanding of the relationship between accountability designations and outcome changes. Evidence of this connection is critical to determining whether the accountability system is supporting the state's theory of action (TOA). We recommend states collect evidence supporting the claim that the accountability system incentivizes behavioral changes and continuous improvement efforts.

Evaluating the Utility and Impact of Accountability and Support Systems

In order to collect validity evidence related to the impact and utility of the accountability and support system, we need to establish a link between accountability data, local behaviors, and outcome improvement. Well-defined systems of improvement will likely be informed by successful past practice and literature reviews pointing to relevant evidence-based practices. However, it is critical to understand the degree to which practices and strategies are appropriate and useful in local contexts. One way this might be accomplished is to leverage existence proofs of success, learn from those context-specific instances, and monitor the application of those strategies to high needs schools aligned to SEA identification plans. The lessons learned from these cases can then be applied to other situations as the accountability system matures and SEAs identify low performing schools in the next cycle.

Taken together, these three sections represent a flow of data and decision making that can be used as a feedback loop to make system improvements. As SEAs learn how their accountability systems function in conjunction with their support systems, it will become evident where adjustments are appropriate or necessary. Those adjustments may be specific to individual indicators, how indicators are combined or weighted, how school identification triggers improvement behaviors or outcome expectations, or how evidence-based strategies are implemented or evaluated.

EVALUATING ACCOUNTABILITY SYSTEMS

While state accountability systems are often subject to a one-time review and approval, accountability systems are not static in nature. As school and district needs change, outcomes improve, and new data are available, accountability systems must be evaluated, revisited, and (as necessary) improved regularly. That is not to say that accountability systems should change regularly—that is akin to changing rules mid-game. Rather, the accountability system and its rules should be adjusted as necessary to align with the system's intended signals of and expectations for continuous improvement. These signals and expectations are a key aspect of a well-articulated theory of action, which should inform any evaluation.

Theories of action articulate how the accountability system is intended to function in order to bring about the desired outcomes. As these theories are tested through system implementation, they

should be revisited frequently to confirm that the underlying assumptions hold. At the highest level, theories of action can be conceptualized at the on-set of accountability system design using the following process:[1]

1. Clearly describe the goals of the accountability system.
2. Articulate the purposes and intended uses of the accountability system results.
3. Define the specific intended outcomes of the system.
4. Lay out the mediating outcomes or intermediate steps necessary to achieve the ultimate outcome(s).
5. Create an initial "high-level" (large grain size) theory of action as a first step to mapping out the components.
6. Build off the "high-level" theory of action and add enough details to articulate how these major components relate to the minor components.
7. "Zoom-in" on several key components of the theory of action to add the detail necessary to support the accountability design and the validity evaluation.
8. Complete the chain of logic by articulating the underlying assumptions which must hold in order for the system to function as intended.

By detailing the goals, purposes, components, processes, and underlying assumptions of an accountability system, policymakers and practitioners will have a strong foundation for evaluation. Evaluation can begin with checking the assumptions to ensure they are being met (e.g., the system of aggregate indicators is reliable enough to support high-stakes decision making about schools). Evaluation can then move to examining the impact the accountability system has on the intended behaviors of those individuals using accountability information (e.g., behaviors of school personnel). Lastly, and perhaps the most importantly, we can evaluate the validity of an accountability system by monitoring progress on the outcomes it is intended to affect (e.g., student achievement).[2] While there are many ways we can evaluate accountability systems, this paper will focus on the concepts of (1) reliability, (2) accountability processes and intended behaviors, and (3) their utility and impact. These three areas were selected because they represent a natural sequence for data and behavior that can be approached as a feedback loop. As lessons are learned, adjustments can be made that may change how reliability is demonstrated in the accountability system or how decision points and their antecedents are linked.

1 Marion, S. M., Lyons, S., D'Brot, J. (2016). *Developing a theory of action to support high quality accountability system design*. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved November 15, 2017, from http://www.nciea.org/publication_PDFs/ESSA%20Accountability%20Design%20Considerations_021916.pdf.

2 Readers are also encouraged to review Hall, Domaleski, Russell, and Pinsonneault's paper on supporting accountability evaluation using a theory of action.

EVALUATING THE RELIABILITY OF ACCOUNTABILITY SCORES AND DESIGNATIONS

School scores have been a subject of interest for local accountability since before the passage of the No Child Left Behind Act of 2001 (NCLB) and subsequently for federal accountability as a result of NCLB, Flexibility Requests from the Elementary and Secondary Education Act (ESEA), and the most recent passage of the Every Student Succeeds Act (ESSA). ESSA maintains the NCLB requirement to administer statewide assessments aligned to state content standards for use in accountability. Sound assessment design, development, and administration require developers and practitioners to collect evidence of fairness, reliability, and validity.[3] Like assessment systems, accountability systems are subject to the same needs. However, how these concepts are operationalized differs depending on whether we are examining the components that comprise an accountability system or the accountability system as a whole.

Reliability of Indicators

Reliability is a concept that is regularly considered in assessment and research. Reliability can be conceptualized as an estimate of the consistency of scores on an assessment while taking into account a variety of ways to quantify error (e.g., indices of internal consistency and structure, mean measurement error, sampling error).[4] Measurement error can be defined in many ways, but is often used to represent the difference between an observed score (e.g., scale score or proficiency claim) and a true score (i.e., the true ability of a student using an idealized set of assessment tools and conditions).[5] When considering measures in an accountability system, reliability includes at least two areas of focus: (1) the consistency of the scores and (2) the consistency of the included population.

It is important to note that these concepts are regularly used when evaluating assessment data. When accountability systems use non-assessment data (e.g., climate surveys, attendance data, grades in college), it becomes critical to consider additional criteria because of (1) softer constraints around standardization and (2) the potential misalignment with the intended use of the data and its application in accountability. Further, the development of tools to collect these data may have fewer resources than data traditionally used for accountability (e.g., high-stakes assessment data), increasing its susceptibility to sources of error. Therefore, we recommend examining data quality criteria related to the unit of analysis, the level of inference, potential corruptibility, and the level of data burden. Readers are invited to review Marion and Lyons' (2016) examination of these concepts specific to identifying measures for the indicator of school quality and student success (SQSS).[6]

3 AERA, APA, & NCME, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA

4 AERA, APA, & NCME, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

5 Brennan, R. L. (2006). Perspectives on the Evolution and Future of Educational Measurement. In. R. L. Brennan (Ed.), *Educational Measurement* (pp. 1-16). Westport, CT: Praeger.

6 Marion, S. & Lyons, S. (2016). *In search of unicorns: Considering the "5th Indicator" in ESSA accountability.* Dover, NH: National Center for the Improvement of Education.

Accountability designers must remember to identify the risks or sources of uncertainty associated with each measure included in the system so that, to the extent possible, their impact on school scores/ratings can be mitigated. For example, sources of measurement error for an assessment-based indicator may result from inconsistency of the content and standards assessed across time and grades, equating considerations, differential motivation, administration factors, and scoring changes. Similarly, it is important for designers to consider characteristics of the population that may influence the consistency of school or state-level outcomes over time. For example, how transient is the population? How inclusive of students is the measure across the state? What types of idiosyncrasies are demonstrated by the population?

After risks or sources of uncertainty are identified, policymakers and practitioners can begin to collect evidence and empirically address questions about the reliability of measures in each indicator. While likely the easiest types of questions to address, the reliabilities of the measures have a direct impact on the reliability of the system (e.g., overall school scores, accountability system, or composite scores) when combined. Therefore, designers must carefully consider the impact that any low reliability component will have on the system's overall reliability.

It is important to note that the components of the system are being used to inform accountability determinations at a school-level. Therefore, the reliability of the component scores or ratings should be examined at the school-level, rather than at the student-level.[7] Once evidence supporting the reliability of measures is collected and examined, policymakers and practitioners can begin to consider the potential impact of and the characteristics of and interactions among indicators.

Interactions Among Indicators

Understanding how and to what degree indicators demonstrate expected characteristics and relationships is necessary to interpret the reliability of school accountability results. This includes considering, for example, whether the variances are similar across measures and whether the relationships between measures are as expected. There are many statistical techniques that can be used to evaluate these factors and consider how they may influence estimates of reliability, several of which are discussed in this section.

Like evaluating the reliability of indicators and accountability scores, it is important to work from the part to the whole when understanding interactions. As an example, many index-based systems rely on policy-specified weights. Assuming the measures exhibit similar score ranges, a brief comparison of the variance of each measure can indicate whether certain measures will exhibit more of an impact on the overall score than others. Average daily attendance is a common example of a measure that tends to have little relative impact on differentiation due to lack of variability especially when compared to something like the rate of students with at least 90 percent attendance or rates of disciplinary referrals in a school. When significant differences

7 Akin to the idea of *unit of analysis*. See Hox, J. (2002). *Multilevel analysis: Techniques and applications* (pp.2-4). Mahwah, NJ: Erlbaum.

exist between policy weights and observed statistical weights,[8] it may be necessary to adjust calculations to honor design principles but account for idiosyncrasies in data.[9] If applied to the previous example of attendance, it is likely that most schools in a state will be awarded a similar number of points (or similar ratings). Thus, any identified policy weight for attendance will be functionally reduced because of the lack of variability compared to any other likely indicator in the system.

These adjustments will likely reduce the transparency and clarity of the overall school score or rating for stakeholders. Consequently, it is important to consider these factors during system design and inform stakeholders along the way to both prepare them for the adjustment and potentially use them as a soundboard for any messaging the SEA may use in communications. Some example analyses and considerations for each are described table below.

Table 1. Sample analyses and considerations for understanding the relationships among indicators

Analysis	Consideration
Correlation: Generically, the relationship or statistical association between a pair of variables.	Depending on whether the system's theory of action is intended to deepen or replicate facets of the accountability system (e.g., drop-out rates and early warning indicators; proficiency-based tests and competency-based performance assessments) higher correlations may be desired. If the theory of action seeks to broaden the measurement of school quality, redundancy of information may be problematic.
Regression: A statistical process for understanding the relationships among variables when conditioning on a dependent variable (e.g., an index score or category).	Regression analyses can be applied to overall accountability index scores (i.e., ordinary least squares regression) or to categories of schools when decision matrices are applied (i.e., logistic regression). Regression analysis can help designers understand the relative predictive power each indicator may have, along with any pooled predictive power on overall scores or designations.[10]
Factor analysis: A statistical analysis that either explores or confirms how indicators relate to one another.	While related to regression, factor analysis can help designers understand whether the indicators relate similarly to the "overall" accountability score or category. Like regression analyses, it requires designers to articulate whether indicators should be related or different. For example, a state may choose to use multiple measures for the SQSS indicator. While separate from the achievement or growth indicators, science assessment data should relate more to academic information than something like behavioral or engagement data. This analysis can help confirm assumptions of planned similarity or dissimilarity.

8 Readers are encouraged to read Di Carlos' (2012) blog post on weighting and the following two references for more information: Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd. Ed.). New York: Springer-Verlag; Wang, M. and Stanley, J. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

9 A state may initially define the policy weights of a set of indicators (e.g., achievement, growth, ELP, and attendance are set to weights of 40%, 40%, 10% and 10%, respectively). Designers will then need to understand the variability of each indicator and how they compare to the variability of the indicators when combined. This can be done, for example, by first multiplying the policy weight by the standard deviation of each indicator (i.e., component variance). The effective component weight can then be calculated by dividing indicator's component variance by the sum of all indicator component variances. This will allow designers to compare how much influence any given indicator will have on the overall score when compared to the policy-specified weights.

10 Nimon, K. & Reio, T.G. (2011). Regression commonality analysis: A technique for quantitative theory building. *Human Resource Development Review*, 10, DOI: 10.1177/1534484311411077.

Analysis	Consideration
Structural equation modeling: A statistical analysis used to understand the structural relationship, or path, between components.[11]	Structural equation modeling is model-based, so it can provide concrete confirmatory information that informs the relationship among indicators and outcome scores or designation of schools, but it can quickly become much more complex to calculate and interpret. This can be particularly useful if there are multiple measures within indicators that are assumed to contribute to the same underlying construct—whether as a group of indicators or the school score/ designation as a whole.
K-means clustering: An analysis that clusters observations (e.g., school scores) into groups based on the closeness of each score to the average score of each group.[12]	This kind of analysis is used to examine how schools are grouped based on the indicators, rather than on cut points or decision rules. That is, how do schools naturally cluster based on their outcome scores? This analysis can help designers understand what significant predictors of cluster (or group) membership emerge. For example, demographic characteristics may be the strongest predictors of school scores or designations. If the theory of action specifies that the accountability system should be sensitive to behaviors or interventions, this may be problematic.[13]

The table above is intended to provide a sample of analyses and how they may be used to help evaluate how indicators interact to better evaluate the accountability system. Further, we recognize that demographics are often a driver of measure variability and will affect the information presented in the analyses listed above. Accountability system designers should consider how their systems detect evidence of improved practice through the equity and fairness of the measures selected. For example, are all students included in certain measures, do changes in n-size incentivize certain enrollment practices, or do indicators prioritize data of lower performing students rather than students of a certain racial or ethnic group (see Domaleski & Perie, [2012][14] for more detail on approaches to emphasize equity in accountability systems)? Understanding the characteristics of each component or indicator can better help us identify the interactions among them. Once explored and documented, we can use our understanding of interactions to help us evaluate the reliability of the school's scores or designations.

Reliability of School Scores or Designations

States and districts have historically struggled to quantify the reliability of school scores, consistency of school classifications, and stability of scores over time. Reliability and error are

11 Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.

12 Forgy, E. W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications." *Biometrics*. 21: 768–769.

13 Demographic characteristics as the most significant predictor could signal a potential flaw in the design of the accountability system that violates notions of fairness and equity. There may be a need to identify measures that are less related to characteristics of the population. That is, identify measures that differentiate school scores by instruction, interventions, and support structures rather than school demographics. Alternatively, the system may seek to be sensitive to demographic characteristics of schools *at the time of design*. Over time, however, changes in indicators or scores should reflect changes in practice and not changes in demographic characteristics.

14 Domaleski, C. & Perie, M. (2012). *Promoting equity in state education accountability systems*. Dover, NH: National Center for the Improvement of Educational Assessment. http://www.nciea.org/sites/default/files/publications/Promoting%20Equity%20CSDMP110712.pdf

common concepts that measurement professionals and SEA officials use when examining the certainty in assessment scores. While the reliability of assessments and other measures raises questions about consistency over time, the same questions aren't sufficient in an accountability context. It may be tempting to think about the reliability of school scores or designations as the consistency of school rankings over multiple years. However, reliability would be more appropriately conceptualized as the likelihood of misclassifying a school.[15]

While policymakers and practitioners might use the relative ranking of schools over time as a proxy for classification consistency, it is more important to evaluate the reliability of a school score or designation as a function of its "correctness" and consistency coupled with the subsequent behavioral expectations associated with a school's score or designation. That is, can we say with sufficient certainty that a school received an appropriate score or designation and that the expected behaviors attached to that designation are also appropriate? While the behaviors associated with designations are not necessarily an element of the reliability of a school's rating, the school's rating has a direct effect on the behaviors schools are expected to implement. Therefore, it is important to answer the question of appropriateness. To do so, we must first understand what factors contribute to the reliability of accountability scores or designations.

Although consistency is raised frequently in this section, we should expect a degree of variation in school scores from year to year. That variation can be attributed to both measurement error (i.e., the error associated with the assessments or tools comprising the components of the accountability system) and sampling error (i.e., the error associated with a different group of students being tested or used for accountability analyses each year).[16] In addition to the variation based on the measurement tool, there is also the natural variance that may emerge because of personnel changes or instructional behaviors in a school.

Sampling error may be the most relevant to consider because of the way in which school-based intervention effects and school composition interact. For example, the type and effectiveness of interventions and instruction vary in schools from year to year. This in turn affects the magnitude and direction of outcome changes from year to year. However, it is unknown whether the observed improvement (or decline) is based solely on the instruction and interventions that have taken place or whether the changes in outcomes are a function of the sample of students and teachers. If we have done our due diligence with quantifying and evaluating the reliability of indicators, we can begin to examine the reliability of the school scores or designations to better understand the impact of sampling in the accountability system.

Hill and DePascale (2002)[17] offer several approaches to help SEAs understand the extent to which measurement and sampling errors may affect an accountability score or system. The authors

15 Hill, R. & DePascale, C. A. (2003). *Adequate yearly progress under NCLB: Reliability considerations.* Paper presented at the 2003 Annual Meeting of the National Council on Measurement in Education: Chicago, IL.

16 See Shavelson, R. J. & Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 277-301.

17 Hill, R. & DePascale, C. (2002). *Determining the reliability of school scores.* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved November 15, 2017, from http://www.nciea.org/sites/default/files/publications/CCSSO02_Reliability_RHCD03.pdf

provide four practical methods and examples for NCLB that are applicable to accountability design in general. The methods, which are outside the scope of this paper, prioritize the ability of a system to estimate the probability that an accountability system will accurately classify schools. If there is a high probability of correct classification using multiple methods, there is increased evidence that the accountability system is functioning reliably (i.e., accurately classifying schools).

Examining the indicator/component reliability, interactions among indicators, and reliability of school scores is a critical step to understanding the utility and impact of an accountability system. Before discussing the utility and impact of a system, we present how these three facets of reliability are related from the part to the whole (i.e., bottom to top) in the figure below.

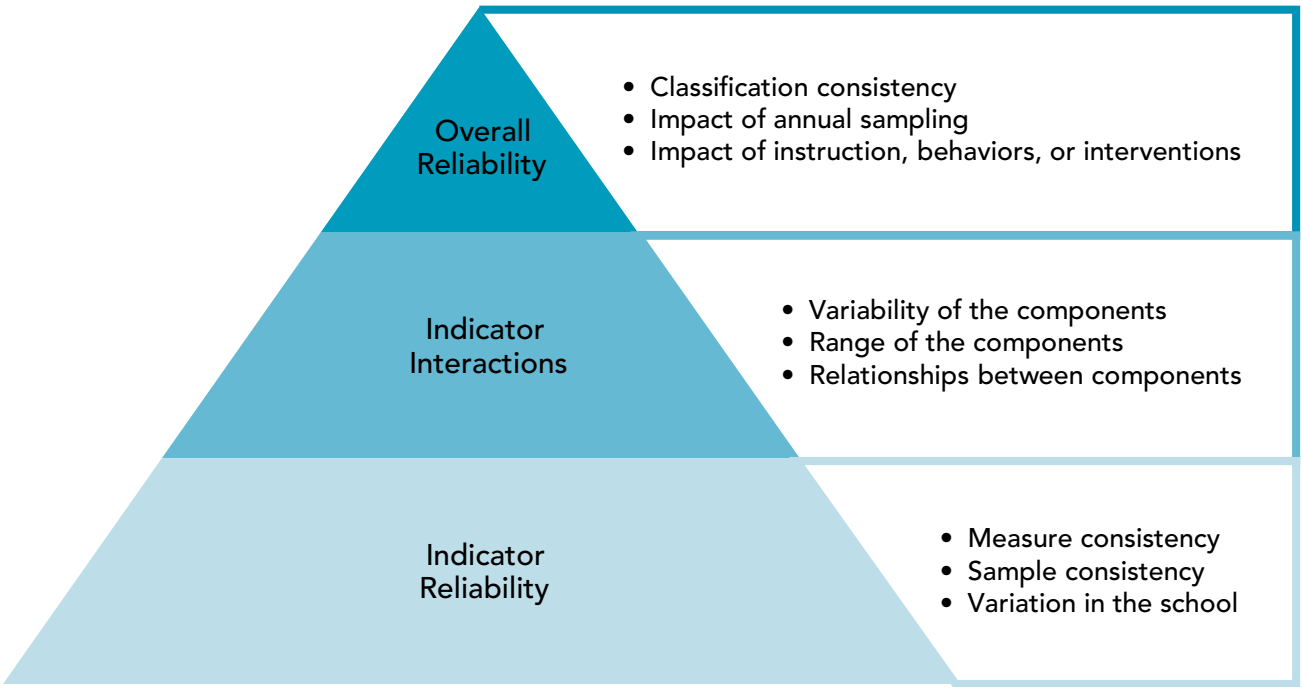

Figure 1. Factors Influencing Facets of Reliability

EVALUATING ACCOUNTABILITY SYSTEM DECISIONS AND INTENDED OUTCOMES

Ensuring that the mechanics of calculating school designations function reliably and support the theory of action is only part of evaluating an accountability system. While school designations are outcomes of the accountability identification, they still only focus on the process portion of the accountability system as a whole. That is, classifying data and identifying schools is the initial step in specifying the behavioral and support expectations for schools. Without documenting how the indicators and system align to the theory of action and operate as intended, we cannot evaluate the utility and impact of the accountability system.

In order to eventually evaluate the utility and impact of an accountability system (described in the next major section of this paper), there is a need to first verify the connection between school

designations and the subsequent supports, interventions, processes, and procedures schools are expected to adopt and implement. If that connection is present, we can begin collecting evidence that the interventions and procedures implemented by schools are useful and making an impact on long-term student outcomes. For example, the way in which schools are differentiated should result in need- or context-specific behaviors aimed at continuous improvement. Those behaviors should then result in beneficial changes in outcome data (e.g., overall accountability score/designation or indicator performance).

There are, however, two antecedent conditions that should be evaluated before exploring the link between accountability designations and outcome changes:

- How do dependencies in the accountability system (i.e., precursor requirements or triggers in the system and their subsequent effects or consequences) affect the perceived utility and value of the accountability information?
- How do exit criteria impact the value placed on behaviors and interventions focused on continuous improvement?

Dependencies between different indicators or components of the accountability system can greatly influence the impact and utility an accountability system will have on intended behaviors or practices. For example, if the dependencies between school designations and performances are poorly specified, schools may disregard the information that the accountability system provides. This issue is described in more detail below.

Dependencies within Accountability Systems

One key area of evaluating the intended impact of an accountability system is based on where dependencies in the system exist. Dependencies in systems have a major impact on how schools are differentiated. If schools are not differentiated with meaning, it will be difficult to communicate why educators and administrators should engage in improvement behaviors. An example of this can be found under NCLB where schools faced a "school size" challenge. Typically, the larger a school is, the more subgroups we can identify in the school. The conjunctive nature of NCLB targets provided larger schools with more opportunities to fail.[18] For schools, the dependency between the number of subgroups and exponential number of targets (i.e., 3 targets per subgroup per grade) created unreasonable expectations that may have been equally a result of sample instability and performance (see Hill & Pascale, 2002). The unreasonable number and size of targets eroded confidence in the accountability system, thereby making the information gleaned from the system less valuable.

Under ESSA (and as evidenced by challenges using Annual Measurable Objectives (AMOs) in ESEA Flexibility), states should carefully evaluate how they use long-term goals (LTGs) or measures of interim progress (MIPs) as part of the system's identification. For example, when LTGs or MIPs

18 Under NCLB, schools had to make progress against annual measurable objectives (AMOs) for proficiency, participation rate, and attendance or graduation rate (depending on grade configuration) for each subgroup of students. If schools failed to make progress on any subgroup, they did not make Adequate Yearly Progress.

are part of the decision rules to assign schools as either state-specific designations or TSI/CSI designations, it is necessary to consider how outcome data or targets (e.g., improvement against LTGs or MIPs) affect changes in state-specific designations or exit from TSI/CSI designations. As with some accountability plans under ESEA Flexibility, schools could demonstrate improvement based on changes in behaviors but not to the extent that schools exceeded LTG or MIP targets. SEAs should determine whether these conditions (if present) are appropriate, whether improvement criteria linked to designation changes should be adjusted, or if outreach efforts are necessary to support administrator and educator understanding of the reasoning behind improvement expectations. The following questions may be useful in reflecting on dependencies, specifically with LTGs and MIPs:

- Are school designations or movement across designations dependent on less high-stakes process or outcome indicators, or is progress against LTGs or MIPs required to change designations?
- Do schools have to collect or demonstrate evidence of engaging in more thoughtful improvement practices in order to exit TSI or CSI designations or to move up in state designations?
- Are there expectations for changes in collaborative or improvement practices that supplement necessary changes in outcome data?

These questions highlight the potential benefit of identifying and using lower stakes process information. These processes could contribute to a more comprehensive understanding of why schools are improving and can help validate the impact and perceived utility of the accountability system. Further, this can substantiate assumptions initially defined in a theory of action or support revisions to the links between examining accountability data and the assumed improvement practices that lead to changes in outcomes.

For example, an SEA may identify TSI schools based on a subgroup demonstrating two consecutive years of performing in the bottom 10th percentile of the state. In order for a TSI school to exit, based on criteria developed by the LEA, that school may be required to demonstrate **both** improvement against subgroup-specific performance targets **and** proof of a high-quality set of evidence-based strategies aligned to subgroup- and capacity-specific needs. Because the school has determined that educators struggle with the vertical pacing of their mathematics curricula, they decided to engage with an external partner to better understand how instruction should be connected using more detailed curricula across elementary grades. This change will be coupled with a focused examination of common misconceptions and foundational gaps for the subgroup in question.

The assumption in the SEA's plan is that they believe evaluating the improvement plan and self-monitoring against the plan at the school level will accelerate increases in subgroup performance. This in turn will lead to the school making its subgroup-specific targets. Without the conjunctive process requirement, it may be possible that the school demonstrated the required improvement because of changes in student enrollment or because of one-off

strategies. A comprehensive and aligned plan will likely promote systemic changes in the instruction and curriculum of the school leading to more lasting changes over time. This type of example is described in more detail in the next section.

Examining Exit Criteria

Exit criteria also play a role in how accountability systems and information is perceived, received, and used. While exit criteria are an example of a dependency within an accountability system, they warrant independent examination given the negative consequences that can occur if they are poorly specified. With states around the country experiencing numerous changes in assessment and accountability systems, exit criteria are often developed with little in the way of historical trend data, which would allow us to better understand the impact of common school improvement strategies. Mis-specified exit criteria may result in schools that demonstrate improvement on accountability indicators but do not improve school designations.

This is similar to issues faced by schools and districts under NCLB. With NCLB, educators and administrators (at schools, LEAs, and SEAs) realized that a system that defined all schools as failing was not differentiating success from challenge. Therefore, the system did not provide actionable information which led to looking for valuable information elsewhere. This same risk exists when considering exit criteria for CSI and TSI schools. For example, exit criteria tied to aggressive MIPs will likely result in schools chasing unattainable targets that continuously increase despite potential gains. This may also result in less differentiation over time that will weaken the value of accountability information. One way to evaluate exit criteria is to classify requirements and determine which ones are expected of schools. Three types of exit criteria are described in the table below.

Table 2. Exit criteria types and descriptions

Exit Criteria Type	Description
Process Requirements	Process requirements require demonstrating evidence of improvement practices (e.g., comprehensive needs assessments, root cause analyses, collaborative partnerships with historically improved like-schools, professional development matched to school needs, implementation of evidence-based practices) that schools must engage in depending on their area of need. These requirements prioritize a process-oriented approach to continuous improvement and assume the right behaviors will lead to success.
Outcome Requirements	Outcome requirements require changes in outcomes that reflect improved educational opportunities for students in low performing schools. These requirements prioritize a results-oriented approach to continuous improvement. Outcomes may be observed in accountability-based outcomes or in improvements on upstream (i.e., antecedent) data that will influence accountability data over time (e.g., improvements in proficiency rates over time, sustained higher rates of student growth, increased graduation rates, sustained increases in English language acquisition progress over time).
Combination of Process and Outcomes	Combination requirements specify that process-based practices and changes in behaviors must be coupled with some evidence of improvement in outcomes. There requirements prioritize a more comprehensive view of school improvement and can be particularly rigorous if outcomes are tied to rigorous high-stakes accountability data expectations.

As noted in the table above, evaluating whether outcome data should be based on accountability-specific data or antecedent data that will eventually lead to accountability-specific outcome improvements may have a large impact on whether schools change designations. For example, the lowest and highest performing CSI schools may exhibit very different performance gaps from the 5th percentile cut point. It is reasonable to expect that schools just below the cut may demonstrate small changes in outcomes that satisfy exit criteria. However, states must evaluate whether this exit is a function of the volatility in the data or if it is a product of both behaviors and improvements in performance. Conversely, the lowest performing CSI school may face significant hurdles to close the gap between their current and expected performance. Larger performance gains may be due to substantial changes in behavior, but may be insufficient to reach CSI identification thresholds. Evaluating whether exit criteria should shift to reward progress toward the exit target in conjunction with evidence of sound improvement practices is appropriate can yield positive outcomes and perceptions of the accountability system.

This evaluation of the antecedent indicators should also include an examination of the entity that owns the source of data and the target location or onus of impact. That is, the state, district, or school may be responsible for certain types of data (e.g., statewide assessment data at the state, interim local assessments at the LEA, school improvement plans at the school), but the onus of improvement practices to impact the data will likely start at the school or district. Recognizing the lag time between changes in behavior and movement in outcomes is important to informing the time interval between monitoring sessions or self-evaluations. Furthermore, a clear identification of those data sources that inform longer-term changes in accountability outcomes can help create logical connections between leading and lagging indicators associated with school improvement efforts. This can be used as evidence in school improvement plans. Additionally, collecting this evidence can help identify model conditions other schools and districts may wish to replicate, adopt, or adjust.

Relationships between Outcome Changes and Accountability Designations

ESSA articulates that states have quite a bit of latitude with classifying schools, as long as the two categories of Targeted for Support and Improvement (TSI) and Comprehensive Support and Improvement (CSI) schools are included (Additional Targeted Support and Improvement is also named, but can potentially be a subset of TSI schools). The latitude offered by ESSA affords states with an important opportunity to develop an accountability system that is coherent with a robust theory of action that emphasizes school support instead of only identification or punishment. This coherence should encompass any additional school designations. As accountability systems are implemented, states should check their assumptions for including or excluding state defined categories of schools. If additional categories were included (e.g., A-F letter grades, star ratings, descriptions), how do they interact with TSI and CSI categories? One example of how this may be approached is to compare school categories to school behaviors. Does the differentiation in state-defined categories reflect differentiation in practices or behaviors?

For example, TSI and CSI schools may actually be expected to engage in the same behavior, where TSI schools "target" their behaviors to the struggling student groups in question and CSI schools apply those behaviors "comprehensively." Alternatively, a school that is rated a "C" may not have any major reform expectations, whereas a "D" rated school must engage in school-wide reform efforts to improve their performance. Understanding this relationship will require examining the differences in behaviors at threshold schools (i.e., those schools at and just above/below cut points) through conversations, interviews, surveys, or observations with administrators, school leadership teams, or LEA/SEA staff who work with the schools directly.

Another way to understand the relationship between identification and behavior is to examine the link between CSI and TSI identifications, how they supplement or replace state-defined designations, and how behavioral changes are promoted through focused school improvement expectations. States should evaluate how their TSI and CSI identification decisions interact with state-specific school designations. In some states, the CSI labels function as a standalone designation that supplements the lowest state rating. In others, a CSI label aligns with the lowest state rating. Because TSI identification decisions are a bit more complex and less consistent, many states have used TSI as a supplementary label as well.

In both the TSI and CSI cases, states should engage in outreach efforts to understand whether the public and educators understand the distinction between designations when TSI or CSI are supplemental. How do the designations communicate different expectations? Have they resulted in different behaviors when examining school data, engaging in needs assessments, or with the extent to which root cause analyses are successful? Identifying possible communication risks can help SEAs develop additional resources or strategies to avoid pushback against the accountability system. This is especially true when considering the risk factors or path that takes a school from TSI to CSI status. Highlighting how and why the two labels are connected (or disconnected) will be important to helping the public and educators buy into the accountability system.

EVALUATING THE UTILITY AND IMPACT OF ACCOUNTABILITY SYSTEMS ON CONTINUOUS IMPROVEMENT EFFORTS

Too often, accountability designers are at risk of separating the identification and improvement aspects of accountability. This may be due to the more quantifiable nature of identification and labeling strategies. However, we recommend that accountability designers and policymakers keep the improvement efforts at the forefront of accountability design. That is, the identification system should serve to bring expectations and behaviors associated with continuous improvement to the foreground. In doing so, SEAs operationalize the TOA of their accountability *and* improvement system.

The prior two sections focus on evaluating the mechanics of the system, system components' interrelationships, dependencies, and antecedent markers of success. This focus is intended to help identify confounding variables that may make it difficult to understand the impact and utility of the accountability system on continuous improvement efforts. In this section, we provide

some sample approaches to understanding how the accountability system may be supporting continuous improvement efforts. Before diving in, we first present the following table, which presents the ESSA language related to school improvement efforts with regard to CSI, TSI, and Additional Targeted Support and Improvement (ATSI). Additionally, we provide considerations for evaluating improvement efforts in these types of schools that are discussed in greater detail throughout this section.

Table 3. Commented Language Regarding CSI, TSI, and ATSI Improvement Efforts

ESSA Language*	Considerations for Evaluating Improvement Efforts
(d) School Support and Improvement Activities.— (1) Comprehensive Support and Improvement.— (B) Local Agency Action.— Upon receiving such information from the State, the local educational agency shall, for each school identified by the State and in partnership with stakeholders (including principals and other school leaders, teachers, and parents), locally develop and implement a comprehensive support and improvement plan for the school to improve student outcomes, that— (i) is informed by all indicators described in subsection (c)(4)(B), including student performance against State-determined long-term goals; (ii) includes evidence-based interventions; (iii) is based on a school-level needs assessment; (iv) identifies resource inequities, which may include a review of local educational agency and school level budgeting, to be addressed through implementation of such comprehensive support and improvement plan; (v) is approved by the school, local educational agency, and State educational agency; and (vi) upon approval and implementation, is monitored and periodically reviewed by the State educational agency.	As states work with schools and local agencies, the need for aligning support and improvement activities with school-specific needs is clear. We recommend that states collect evidence of the impact of school and local activities with an eye toward making a link between behaviors and tangible improvements in outcome data. The state requirement to monitor and periodically review the comprehensive support and improvement plan provides a strong rationale for tight local partnership to help promote continuous improvement efforts. Recommendations and considerations for states are discussed in this section.

ESSA Language*	Considerations for Evaluating Improvement Efforts
(d) School Support and Improvement Activities.— (2) Targeted Support and Improvement.— (B) Targeted Support and Improvement Plan.— Each school receiving a notification described in this paragraph, in partnership with stakeholders (including principals and other school leaders, teachers and parents), shall develop and implement a school-level targeted support and improvement plan to improve student outcomes based on the indicators in the statewide accountability system established under subsection (c)(4), for each subgroup of students that was the subject of notification that— (i) is informed by all indicators described in subsection (c)(4)(B), including student performance against long-term goals; (ii) includes evidence-based interventions; (iii) is approved by the local educational agency prior to implementation of such plan; (iv) is monitored, upon submission and implementation, by the local educational agency; and (v) results in additional action following unsuccessful implementation of such plan after a number of years determined by the local educational agency.	Increased flexibility to identify TSI schools makes it more difficult to understand conditions for improvement and success. TSI provides an opportunity to understand how interventions may be successful in various school contexts and configurations. TSI also presents a challenge to understand the role of local processes in improvement efforts. States should leverage existence proofs and test cases to understand local contexts and applicability to other schools. Additionally, the local requirement to monitor school plans is a prime opportunity for the SEA to learn about context-specific conditions that promote or hinder successful implementation of improvement efforts. These concepts are described later in this section.
(d) School Support and Improvement Activities.— (2) Targeted Support and Improvement.— (C) Additional Targeted Support.— A plan described in subparagraph (B) that is developed and implemented in any school receiving a notification under this paragraph from the local educational agency in which any subgroup of students, on its own, would lead to identification under subsection (c)(4)(D)(i)(I) using the State's methodology under subsection (c)(4)(D) shall also identify resource inequities (which may include a review of local educational agency and school level budgeting), to be addressed through implementation of such plan.	ATSI schools are a unique opportunity to categorize schools. Depending on a state's TOA or accountability plan, ATSI schools may be a logical step between TSI and CSI schools, or may serve as an entirely unique signal. Thus, SEAs can examine ATSI identification and improvement expectations to confirm assumptions about differentiation in improvement behaviors as a result of being identified an ATSI school. This school identification also provides states an opportunity to determine whether outcome improvements are based on changes in improvement practices (i.e., adopting widespread improvement efforts for a short time) or is it due to the consistency of improvement behaviors (i.e., sustained and focused improvement efforts).

*Identification and exit criteria are specifically addressed in the first of this two-paper series.[19]

19 Lyons, S., D'Brot, J., Landl, E. (2017). *State systems of identification and support under ESSA: A focus on designing and revising systems of school identification*. Washington, DC: Council of Chief State School Officers.

By clearly identifying the step-by-step decisions and mechanisms in the accountability system, SEAs can greatly clarify a theory of action. This clarification can then be used to identify cases of success (i.e., schools that have demonstrated improvement in the face of subgroup-specific or school-wide underperformance) and establish frameworks to monitor improvement over time. This can be done through the use of both existence proofs and test cases, which are described below.

Identifying Existence Proofs of Success

Often, educational researchers and practitioners search for test cases that prove a point or provide evidence of the success of some program. When evaluating accountability systems, interventions, or programs that are repeatedly reinvented, it may be more valuable to identify an existence proof,[20] or a school that satisfies some condition or phenomenon. In this case, what schools or districts have changed behavior and improved outcomes using recognized evidence-based practices that are being promoted under ESSA? Identifying these exemplars can help clarify the situations for success, strategies that can be changed, context-independent conditions, and key intervention characteristics.

Leveraging exemplars like these to learn about a program is referred to as implementation evaluation, or evaluating a program's implementation as designed while identifying issues that surfaced and adjustments that had to be made.[21] SEAs can adopt this approach to help other schools and districts understand improvement programs and strategies by engaging in the following activities or posing the following types of questions:

1. Identifying an exemplar existence proof;
2. Discovering the original proposed implementation and model;
3. Clarifying the original theory of action and whether resources were available for the model's assumptions;
4. Determining whether threats to implementation were identified before, during, and after the program;
5. Understanding how characteristics of the program or practice differed from the design;
6. Verifying the primary activities. Who at the school or district levels was involved?
7. Uncovering program deviations from the original design. Why were changes made and who discovered the need for those changes?
8. Determining whether the final model differed from the original model in terms of the connection between resources, activities, and outcomes;

20 Linn, R. L. (2003, July). *Accountability: Responsibility and reasonable expectations* (Tech. Rep.). Los Angeles, CA: Center for the Study of Evaluation, CRESST.

21 Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

9. Clearly identifying conditions that are necessary for success. Are there cases where this program or intervention would not be successful or recommended?

10. Determining the feasibility of sustained change in practice. How has this program or intervention changed behavior for the long-term? Does this change continue to yield improvements in outcomes?

As an example of the steps above, consider the case of a large school that is predominantly economically disadvantaged with a diverse student body that includes the following student groups: African American, Asian/Pacific Islander, English Learners, Hispanic/Latino, Low Socio-Economic, and Students with Disabilities. Following the passage of ESSA, the school's low performance would have qualified them for being identified as a CSI school. However, prior to ESSA being passed, the school collaborated with their LEA, partnered with a higher performing school, and adopted many of the data review, root cause analysis, and horizontal- and vertical-team strategies to address the following:

- Data analysis approaches: working in collaboration with a higher-performing school to review various data from both state and LEA-sourced accountability reports to engage in more comprehensive root cause analysis
- Student engagement: pockets of low attendance in conjunction with an LEA-sponsored truancy specialist to increase the capacity of school staff to replicate the specialist's resource use and approach
- Teacher engagement: addressing teacher turnover by implementing professional development that qualifies for continuing education credit and creating a teacher mentorship program
- Instructional improvement: partnership with a high-performing school to better understand how to address the cognitive complexity of the standards and specify learning progressions across grades to support vertically cohesive instruction

During the collaboration effort, the struggling school initially replicated the analysis approaches exactly as implemented by the partnering school. While some improvements were made in the first year, the struggling school recognized that their sources of engagement issues were less generalized across the student body and more widespread from a teacher turnover standpoint. This minimized the effectiveness of instructional improvement support because of the lack of coverage in supporting a critical mass of teachers in the school. However, the data analysis approach helped highlight the need to better focus student attendance issues on a much more focused level. This prompted the increased partnership with the LEA to access a truancy specialist.

This struggling school then demonstrated improvements over the courses of years 2-4 and is no longer in the lowest performing category of the state. Given the demographic and socio-economic diversity of the school, it represents many of the currently low performing schools now identified under the state's ESSA accountability plan. Thus, the state can identify this school as an existence proof and work with the LEA and school to determine how they adopted the

approaches of their partner school and subsequently how and why they made adjustments to their improvement strategies.

While this example is general in nature, we hope it helps readers understand how these activities and questions can help school improvement specialists understand why existence proofs were successful. By carefully selecting schools or districts that meet certain criteria (e.g., demographic composition, geographical location, data attributes), we can begin to create profiles of successful implementation based on the type of intervention or program applied. For newer evidence-based strategies, we can begin establishing a framework for identifying successful implementation conditions, context-independent constraints, context-dependent circumstances, and the role of actors at different levels of implementation (e.g., classroom, school, district, state). This can help inform states how to best make improvements to new cases of implementation.

Identifying and Monitoring Test Cases of Success

If existence proofs are about understanding why strategies or interventions were successful, test cases are about confirming conditions and monitoring the magnitude and direction of change over time. Accountability systems offer relatively crisp distinctions to identify test cases for evidence-based practices and possible improvement. A common outcome of an accountability system is school differentiation, with lower performing schools typically designated lower ratings. Those low-rated schools (or schools on the cusp of low ratings) may be ideal candidates for partnership with the SEA to understand how antecedent and outcome data change as a result of focused improvement efforts. If improvement efforts are implemented in conjunction with LEA or SEA support, it offers an additional opportunity to increase our understanding of the conditions that contribute to the effective implementation of improvement efforts.

Once identified, the SEA can begin to conceptualize how it might partner with LEAs and schools to track detailed changes in practice, behaviors, conditions, and data to document behaviors, implementation conditions, changes to consider, and emergent challenges. This also offers the opportunity to confirm whether school improvement specialists are observing the expected changes in short-term outcomes, lower-stakes data, or antecedent outcome or process measures that are logically linked to high-stakes accountability outcomes. This can then be used to confirm or revise assumptions about what proximal behaviors and data sources are linked to longer-term data that prompt changes in accountability designations.

A key part of evaluating the link between behaviors and outcomes is specifying the link between the immediate behaviors sparked by accountability, the intermediate behaviors associated with instruction and intervention, and the long-term changes in low- and high-stakes outcomes (e.g., data reviews, behavioral changes, changes in proximal data, and changes in distal data).[22] Evaluating this link requires drawing logical connections between the following types of activities:

22 Data reviews include activities like needs assessments and root cause analyses. Behavioral changes include things like evidence-based practices or interventions. Changes in proximal data refer to changes in process data, early warning data, or antecedent data associated with classroom or schoolwide practices or interventions. Changes in distal data refer to changes in outcome data that are used in accountability systems that drive changes in designations or required school improvement activities.

- Improvement behaviors;
- Data analysis sources and procedures;
- Improvements or changes in the capacity of school personnel;
- Collaborations or partnerships within or outside of the school or district;
- Revisions, adoptions, or development of resources; and
- Improvements in school or instructional processes and procedures.

Once sufficient detail is available to tell a cohesive story around effort and improvement, the school, LEA, and/or SEA can articulate and disseminate their findings and recommendations to others. This can help schools and states make the necessary connections and evaluate their own ability to support ongoing improvement efforts. A way to consider how these logical connections might look is displayed in the figure below.

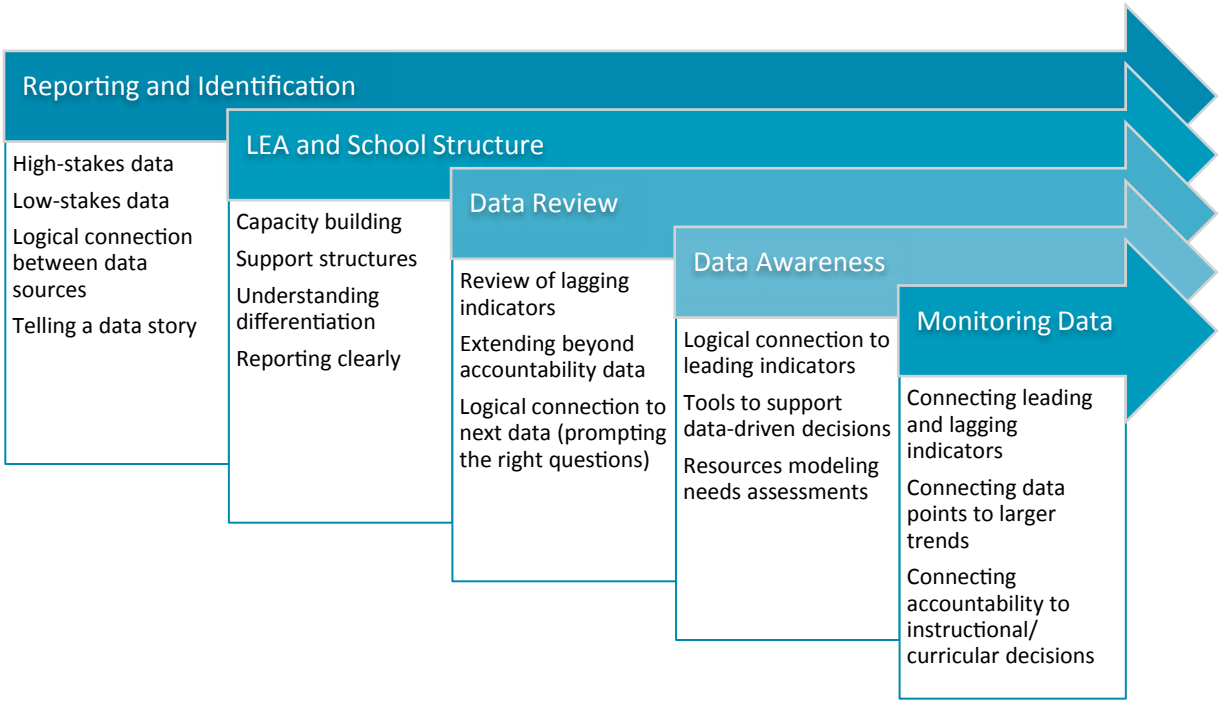

Figure 2. Supporting Continuous Improvement by Connecting Accountability to Local Behavior

The figure above forwards one possible way of linking accountability data to localized improvement efforts. As indicated by the large arrows, there are logical top-down steps that begin with the state (i.e., accountability reporting and identification), include the LEA, and are eventually focused at the school. This may also be considered a detailed sub-section of an accountability system's TOA that is intended to highlight how accountability data should help inform local improvement efforts using data beyond the scope of accountability reporting but well within the scope of strong continuous improvement efforts. Because so much of the traction in school improvement is dependent on LEA- or school-specific information and practice, SEAs may struggle in supporting schools and LEAs if there is insufficient capacity to do so. By engaging in self-evaluation, SEAs

can better support local improvement efforts, better monitor successful improvement efforts, and better convey stories of success with enough fidelity to make them informative. We describe a self-evaluation approach in the next section.

ONGOING SELF-EVALUATION FOR MONITORING ACCOUNTABILITY

The evaluation activities recommended in this paper require time, space, and personnel. In order to engage in this work well, it is important for SEAs to self-evaluate their own capacity and determine what adjustments are needed. There are numerous types of capacities SEAs, LEAs, and schools should consider when implementing a large-scale or comprehensive evaluation (i.e., accountability system evaluation). These include human, organizational, structural, and material capacities. These capacities,[23] which are explained in more detail below, can help an agency understand their ability to evaluate their own accountability system. They are applicable to many reform efforts and can be modified accordingly. These SEA capacities include

1. *Human capacity*:[24] The knowledge, skills, will, and view of self, key stakeholders and those who will be part of the school improvement process. They will have direct or indirect roles that help increase stakeholders' capacity to use data from improvement efforts effectively.
2. *Organizational capacity*:[25] The interactions, relationships, and communications among individuals in the system that shape culture regarding data review, behavioral change, and monitoring. They set the tone for collaboration.
3. *Structural capacity*:[26] The elements within the system that exist independently of the individuals involved. These include features like policies, procedures, and formalized practices of a system.
4. *Material capacity*:[27] The fiscal and staffing resources and other material supports to promote effective use of the system. This includes in-kind time, meeting space, technological capabilities, training documentation, and transportation/travel available to support the dissemination of or professional development in improvement efforts and activities.

23 Century, J. R. (1999). *Determining capacity within systemic educational reform*. Paper presented at Annual Meeting of the American Educational Research Association. Montreal, Quebec, Canada.

24 High-priority stakeholders who are expected to use data from the system for decision making; the communications to high-priority stakeholders regarding the resources available; and the role-specific training for agency staff and high-priority stakeholders on how to appropriately interpret data from the accountability system.

25 Interactions within the SEA to deepen partnerships, relationships, and communications among stakeholders and agency staff; the ways stakeholders should be engaged from planning through implementation; communications by and for agency staff and intended users; facilitation of project management implementation; and identification and mitigation strategies to organizational barriers for effective data use and school improvement efforts.

26 Policies, processes, and protocols developed by the SEA for successful data use; sustainability through the policies and practices that result from implementing work plans; processes for ongoing review and revision of work plans; and identification and mitigation strategies for the structural barriers to effective data use and school improvement efforts.

27 Tools, reports, and supporting documentation necessary for data use and school improvement efforts; enhancements to the system's technical infrastructure; training plan(s) and materials to support data use and school improvement efforts; travel, conferences, and/or professional development opportunities to increase other capacities; and needs related to management and coordination of vendors and products.

SEAs should consider their areas of strength and weakness with respect to each type of capacity. For example, a department may have very knowledgeable staff (i.e., high human capacity) with strong inter-departmental relationships (i.e., high organizational capacity), but none of the policies or procedures necessary to monitor improvement at the local level (i.e., low structural capacity). Until that is addressed, it is unlikely that documentation of needs assessments or resources modeling effective school improvement practices (i.e., material capacity) will be developed. An analysis of internal capacity can help clarify the opportunities or threats to successfully monitoring the implementation and use of an accountability system and inform the degree to which partnerships need to be cultivated to support accountability evaluation.

In the case of accountability, it is relevant to discuss the challenges associated with organizational and structural capacity. In most SEAs, partnerships (i.e., organizational capacity) and processes or procedures (i.e., structural capacity) to support a reform effort are typically internal affairs. To adequately evaluate an accountability model, the SEA must understand how it interacts externally with LEAs and schools. In addition, SEAs must leverage partnerships with local agencies and schools to fully understand how accountability data are being used, whether behaviors have changed, and the degree to which those changes yield improvement in student outcomes.

SUMMARY

The purpose of this paper is to provide guidance and recommendations for how SEAs can approach evaluating their accountability systems. This goes beyond simply evaluating the identification process, but also includes considerations for linking accountability evaluation to behavior change, outcome monitoring, and implementation evaluation. Specifically, policymakers, practitioners, and accountability designers should attend to the reliability of accountability scores and designations, the utility and impact of accountability systems, and the link between behaviors and outcome improvement driven by school identification and accountability.

When evaluating the reliability of accountability scores and performance designations (including identification for support and improvement) it is important to consider how measurement and sampling interact to affect school-level reliability estimates of the measures comprising the indicators. Furthermore, we should conceptualize reliability in accountability designations and scores as classification consistency in the most appropriate category. We can support these two efforts by ensuring we have a clear understanding of how indicators interact and how schools are grouped, which can be accomplished by applying various calculations to observed accountability data and results.

When evaluating the utility and impact of accountability systems, it is important to first understand the dependencies introduced in an accountability system by evaluating accountability decisions and intended behaviors. These could be obvious design decisions or possible unintended effects based on improvement expectations for schools. SEAs should also attend to how exit criteria are defined, how they relate to observed changes over time, and how more proximal information is

linked to distal information typically used as outcomes in accountability systems (e.g., proficiency, graduation rates, and other student or school outcomes). Additionally, understanding the relationship between outcome changes and accountability designations is critical to understanding whether the accountability system is supporting the state's theory of action. What evidence do we have that behavioral changes are spurred by the accountability system and how do we make changes to incentivize continuous improvement efforts?

Finally, when evaluating the impact and utility of accountability systems on continuous improvement efforts, we can begin to validate the impact of the accountability system by collecting evidence of change and success over time. In addition to reviewing literature detailing successful past practices, SEAs can begin by identifying existence proofs of success that fit profiles of schools that are similar to other high needs or low performing schools. These schools offer an opportunity to understand contextual conditions that contribute to their improvement. Experiences can then be applied to test cases where monitoring would likely occur independent of any evaluation or close examination of school climate, conditions, or environment. As a key part of this ongoing monitoring, SEAs should engage in continuous self-evaluation to determine whether there is sufficient capacity to monitor, understand, synthesize, and apply lessons learned to other contexts or situations. By developing a comprehensive evaluation strategy for both the identification and utility associated with an accountability system, SEAs can forge or leverage internal partnerships to increase the impact and efficiency of educational improvement efforts.

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072