Accountability Validation: Evidence as a Function of Design

Juan D'Brot, Senior Associate
National Center for the Improvement of Educational Assessment

Paper in support of the Center's *Reidy Interactive Lecture Series: The Center at 20*

Introduction[1]

The passage of the Every Student Succeeds Act marked the beginning of a new development cycle for accountability systems. This legislation presents State Education Agencies (SEAs) with the opportunity to redesign accountability systems that have been part of the *Elementary and Secondary Education Act* and its reauthorizations under the *Improving Americas School Act*, *No Child Left Behind*, and *State Flexibility from ESEA* (aka Waivers). Specifically, SEAs can reinforce the sometimes unclear connection between accountability and school improvement systems and more fully address the need for ongoing continuous improvement. Continuous improvement is a focus of the accountability principles set forth by CCSSO (2011) on behalf of states, and it provides additional possibilities for strengthening the coherence of accountability and improvement systems with an SEA's larger priorities and theories of action.

Beyond school ratings, state accountability systems are somewhat abstract from the vantage point of schools and classrooms. SEAs tend to leverage accountability systems to incentivize behaviors that improve student outcomes and facilitate equitable access to high-quality educational opportunities. However, there often is a gap between the intended impact of an accountability system and how behaviors actually change. While a strong system design arguably mitigates against such a gap, there nonetheless are many development activities to which designers should attend. Figure 1 illustrates the continuous nature of an accountability and improvement system, which is further described later in this paper.

[1] This paper is based on the recent publication by D'Brot, Keng, & Landl (2018), which can be found here.

Figure 1. The Accountability and Improvement Cycle

Accountability and improvement systems are multi-stepped and multi-layered, which creates a series of dependencies that require confidence in preceding system decisions to support the validity and accuracy of subsequent decisions. Thus, early evidence of the system working as intended (e.g., design decisions, business rules, and school performance expectations) can make it easier to confirm that the SEA has identified Local Education Agencies (LEAs) and schools in greatest need of support and services. In this paper, I provide an overview of why we should consider developing validity arguments—the process of collecting and interpreting evidence to support decisions—for our accountability and improvement systems. In doing so, I also consider issues that surface as we address the burden of evidence faced when designing accountability systems.

First, I argue why validity arguments are necessary for accountability and improvement systems and how the school identification- and service delivery-focus of these systems should be expanded. Second, I briefly describe the need for SEAs to evaluate each activity along the school identification and service delivery stages (see Figures 2 and 3 below). Third, I present a framework (D'Brot, 2018) that can be used to evaluate intended outcomes associated with activities within the stages presented in Figure 2. Fourth, I provide a brief conclusion summarizing the intent of this paper.

Validity Arguments for Accountability and Improvement Systems

From a measurement perspective, validity refers to the ability of an instrument to measure what it purports to measure—however, validity is dependent on a comprehensive set of coherent evidence (AERA, APA, & NCME, 2014). State accountability systems can be thought of as a measurement instrument that helps the public understand the degree to which schools and districts meet the state's educational goals and priorities (see Keng, D'Brot, & Landl, 2018). Therefore, establishing a validity argument for accountability entails identifying and connecting the pieces of evidence so SEAs can be confident that the schools' ratings are accurate, fair, and valid. A validity argument also allows SEAs (and LEAs) to appropriately support schools that are struggling and recognize schools that are excelling.

Accountability Systems Stages

Like any complex system, an accountability system is based on a series of dependencies. The soundness of any one decision is contingent on the soundness of each preceding decision. Documenting these decisions, and compiling evidence supporting outcomes for each decision, helps SEAs make a validity argument for their accountability and improvement systems. However, SEAs must first recognize the important decision points at each step, such as defining a system vision, determining policy weights, or setting performance standards for the accountability and improvement system as shown in the figure below. This can help SEAs identify the most relevant evidence for each decision point. Collecting this evidence can help instill stakeholder confidence in design decisions, system processes, school performance expectations, and the delivery of services and support. Despite these system complexities, there are three general categories of accountability identification activities: design, development, and implementation. The following figure outlines these three categories as they apply to accountability and improvement systems and how these systems identify schools.

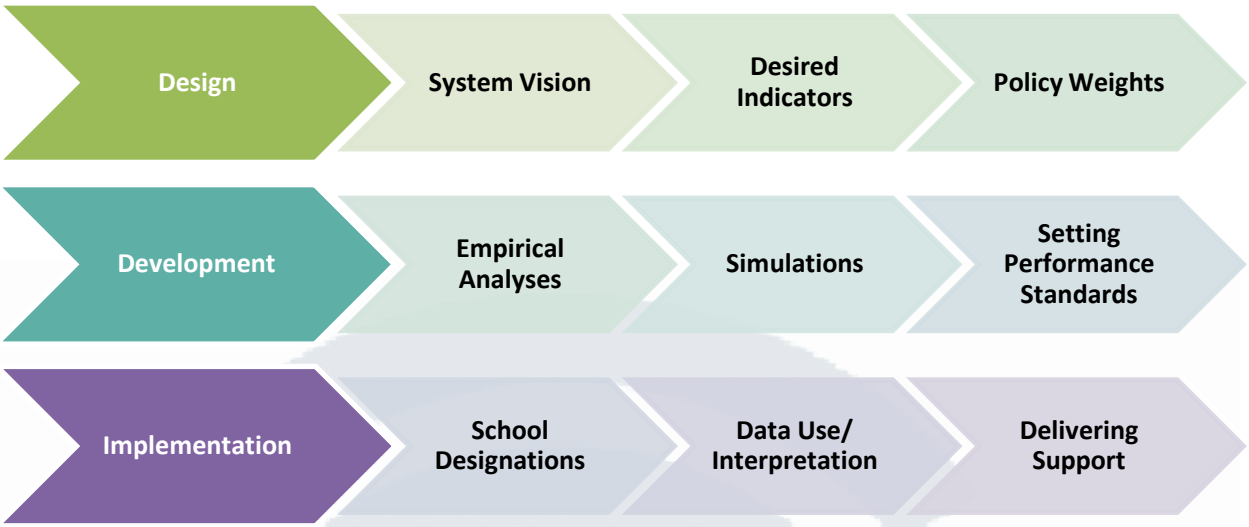

Figure 2. Accountability Design, Development, and Implementation Stages.

The **design** stage includes refining the system's overall vision (e.g., policy priorities, educational system goals, role of accountability), specifying indicators based on the intended outcomes (e.g., growth and achievement, college readiness vs. career readiness, engagement), and determining policy weights to capture SEA values and priorities (e.g., growth should be equally weighted with achievement).

The **development** stage includes evaluating the indicator measures and relationships among them (e.g., descriptive and inferential analyses, qualitative reviews of data and processes), identifying potential data gaps or capacity concerns through the use of simulations (e.g., projections, historical data examinations, mock accountability runs), and specifying performance expectations over time by setting defensible performance standards.

The **implementation** stage includes supporting the determination and release of school designations; helping stakeholders access, use, and interpret accountability data; and assisting the SEA and LEAs as they deliver supports and interventions to schools. These activities inform local inquiries and information use.

School Improvement Systems Stages

The next set of activities pertains to statewide systems of support and the conditions needed to facilitate improvement activities. These activities are alluded to in Figure 2 above, but should be examined in more detail. These activities are presented below in Figure 3 and focus on the delivery of appropriate support services, given the identification decisions made.

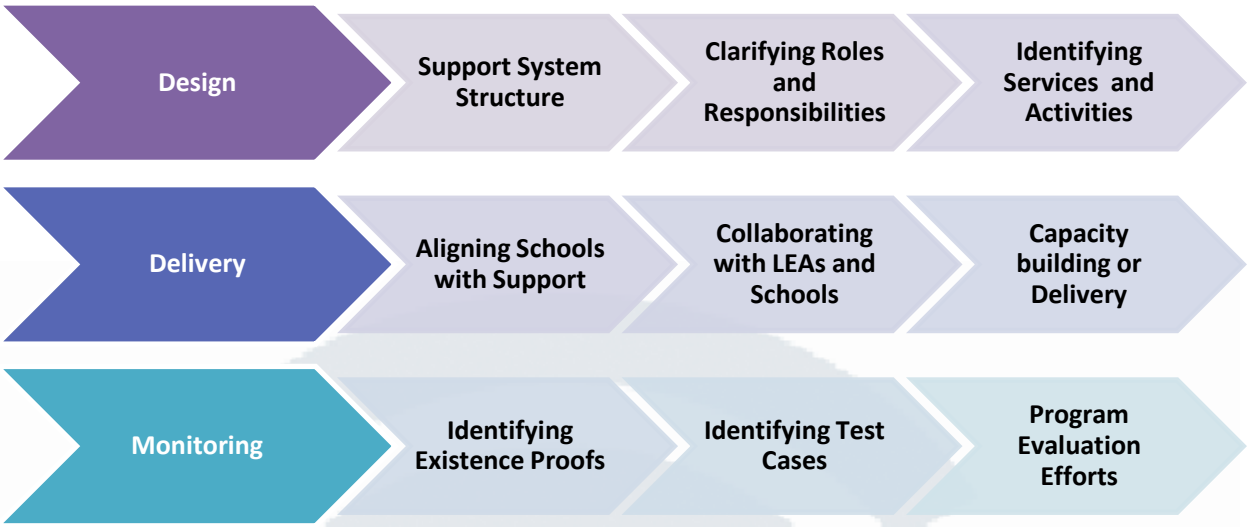

Figure 3. Support System Design, Delivery, and Monitoring Stages.

When compared with the previous figure, Figure 3 shows some similarity in structure but the specific content differs significantly. I would argue that a system of support is broader than simply delivering support. However, an SEA's system of support is typically delivery-focused while the accountability system is more identification-focused, as noted in Figure 2. A key part of establishing a validity argument for accountability and improvement systems includes recognizing the parallels and complementary nature between the identification and delivery focused activities in each.

The **design** stage includes framing the structure of the support system (e.g., information and decision flow within and outside the SEA), clarifying roles and responsibilities (e.g., identifying key SEA capacity builders, determining the role of the LEA, identifying typical school-level contacts), and identifying services and activities (e.g., determining school improvement vision or high-yield evidence-based strategies).

The **delivery** stage includes aligning schools with support based on state and federal designations as appropriate, collaborating with LEAs and schools as specified in the system of support design, and engaging in capacity building or service delivery efforts with LEAs or schools. Through the delivery activities, SEAs can begin compiling progress data and behavior profiles to support efforts in the monitoring stage.

The **monitoring** stage includes identifying cases where schools have successfully implemented strategies in the past (i.e., existence proofs), identifying schools in need of support to track the impact of services provided, and leveraging program evaluation methods to monitor how services improve progress and outcome data. These activities help bolster validity arguments associated with identification decisions local inquiries and information use, which are described in D'Brot, Lyons, and Landl (2017).

While the emphasis of this paper (and the associated session during RILS) is on establishing a validity argument for the accountability and improvement system, the complementary roles of accountability

and improvement are not regularly leveraged by the service-delivery focused nature of the support system. The supports and progress-monitoring associated with an SEA's support system should be used to understand whether the identification system sends the right signals, prompts effective questioning, and elicits the intended behaviors among LEAs and schools. The information gathered from the support and monitoring that states and their partners provide should then be used to confirm identification decisions or refine the system's design.

The remainder of this paper raises critical areas where states are poised to examine practice and collect evidence to support the development of an overall validity argument for their accountability and improvement systems. Immediately following the next section, a framework is presented that can be used to evaluate state accountability design, development, and implementation decisions.

Evaluating Accountability and Improvement Systems

The development of accountability and improvement systems begins with specifying a theory of action, which explains how the accountability system will effect desired outcomes. At a high level, the flow of information across major components of a theory of action (e.g., reporting, identification, consequence and reward provision, monitoring) can be examined and confirmed. At a more detailed level, states can focus on the major stages (and activities) to determine whether assumptions hold true. Each connection between stages and activities (e.g., policy weights to empirical analyses; empirical analyses to simulations; simulations to performance standard setting) offer information hand-offs that should be used to clarify the claims being made at the start of each stage and more detailed activity.

Tested through system implementation, the theory of action should be revisited to confirm that the underlying assumptions are tenable and the processes are being enacted as intended. Further, specifying the associated outcomes is critical to utilizing an evaluative framework (D'Brot, 2018) presented in the next section. Possible outcomes associated with accountability systems might include the following:

- Bring the lowest performing students up to proficiency;
- Encourage the academic improvement of ALL students, including those already proficient; or
- Broaden the range of skills students acquire so they are college- or career-ready.

By detailing the goals, purposes, activities, processes, and underlying assumptions of an accountability system, policymakers and practitioners will have a strong foundation for monitoring and evaluation.

Monitoring and Evaluating Accountability Systems and School Identification

The remainder of this paper describes a framework for states to evaluate identification decisions in a systematic way. The framework first presents example claims associated with each activity shown in Figure 1 (e.g., system vision, indicator selection, performance standards). For each claim, it then provides guiding questions to help SEAs clarify the intended purpose, use, and process associated with the claim to determine the assumptions that must be met. These assumptions, in turn, are used to identify sources of information, methods, or analyses that can be employed to collect information to defend the claim. This framework is not intended to be prescriptive, but rather to provide examples of

how states can apply the framework to begin monitoring and evaluating SEA systems—a key piece to establish validity arguments for accountability and improvement systems.

This framework expounds upon the idea that the correct identification of schools is a necessary but insufficient condition for building capacity and delivering support to Local Education Agencies and schools. Systems of accountability, support, and continuous improvement contain a series of feedback loops and information hand-offs that offer opportunities to collect evidence on whether the systems are working as intended. By identifying activities and their relevant evidence throughout the design, development, and implementation of accountability systems, we can begin to develop validity arguments for our accountability and improvement systems. The following framework (D'Brot, 2018) can support systematic monitoring and evaluating of the design, development, and implementation stages of accountability identification. This framework can be applied to the activities in the design, development, and implementation stages of accountability systems (Figure 1).

Within each of these stages and activities, SEAs can widen or narrow what they monitor to expand or limit system claims. Claims are statements or assertions about the accountability system and its impact and will likely differ in granularity depending on the level of focus. While claims will likely differ in granularity depending on the level of focus, they clarify the kinds of questions state leaders could be asking and the types of evidence they can consider collecting and evaluating.

Using a Framework to Monitor and Evaluate Accountability System Efforts

By developing a set of claims associated with accountability and improvement systems, SEAs can begin developing a logic model that identifies relevant assumptions, questions, data considerations, and possible evaluation tactics. These claims can help states establish a systematic approach to monitoring and evaluating their accountability and improvement systems, thus contributing to an overall **validity argument**. However, the validity of the full system rests on the confidence that states have in the validity of each step or stage along the way. This is similar to the point that system reliability is dependent on component reliability (see D'Brot, Lyons, & Landl, 2017). Specifically, a state must first verify the reliability of indicators and components before examining the interactions among them. A state can then examine the reliability of overall school designations.

For each claim, associated assumptions must be confirmed to support the corresponding step in the accountability and improvement system. These assumptions guide practitioners and designers toward sources of information, methods, or analyses to be used in support of the claim. To narrow the information or approaches necessary for developing a validity argument, I present below a series of questions associated with each of five important claims.

Claim 1: Indicators provide fair and accurate information that informs the accountability system as intended.

This first claim highlights the need for states to have confidence in the components leading to school-level designations. Whether designations are summative (e.g., index-based) or descriptive (e.g., dashboard), the school score components or decision rules used to classify schools are contingent on a series of interactions or branching decisions that depend on how measures within each indicator behave. These guiding questions should be answered to determine the most appropriate sources of information or methodological approaches:

- What is the intended purpose of the indicator within the system?
- How will the indicator be used?
- What inference(s) is the indicator intended to support? Is the indicator an observable count of something being directly measured, or is it a proxy for something else?
- How should the indicator relate to the other indicators in the system?
- What characteristics should the indicator display to support its intended purpose?
- What data are required to inform calculations of the indicator?

In answering these questions, states can determine which of the following assumptions, sources, and analyses are most appropriate.

1. Claim to be Supported by Evidence: *Indicators* provide fair and accurate information that informs the accountability system as intended.		
Assumptions	**Research and Evaluation Sources**	**Example Analyses or Data[2]**
• Indicator scores/ratings are reliable. • Data necessary to inform the indicator are accurate and complete. • The indicator relates to other indicators as intended. • The indicator demonstrates characteristics necessary to ensure it is valid and fair. • Data and procedures support the intended interpretation of results (e.g., standard setting, norming).	• Clear connection between policy rationale and indicator • Specification of measures (i.e., data) that constitute the indicator • Quantitative and qualitative examination of reliability, robustness, objectivity of measure • Relationship of measures within and across indicators • Examining business rules and quality assurance procedures	• Descriptive examination of measures (e.g., distribution shape, central tendency, variability) • Reliability evidence where appropriate • Analyses to determine relationships among indicators (e.g., correlation, regression, factor analysis) • Quality assurance checklists and business rules monitoring[3]

[2] Readers are encouraged to review Goldschmidt's (2018) handbook on developing and monitoring English Language Proficiency indicators, found here.
[3] Readers are encouraged to review Keng & D'Brot's (2018) Operations and Quality Control brief for specific suggestions on this approach.

Claim 2: The state's system of meaningful differentiation provides accurate, reliable, and valid information about the relative performance of schools in a manner that reflects the state's values and priorities.

This second claim addresses the degree to which the system of meaningful differentiation[4] is in alignment with the policy goals and priorities of the state. Whether summative or descriptive, school-level designations send an important signal to stakeholders and, further, should provide sufficient separation among schools and their profiles. Answers to these guiding questions help determine the most appropriate sources of information or methodological approaches for confirming how the larger sets of decisions inform the assignment of school decisions:

- Do aggregate data reflect the same differentiation as student-level data?
- Are intentionally different signals (e.g., achievement and growth) sufficiently complementing each other within the system?
- Is there variability for both the overall score/designation and the various indicators?
- Have simulations been conducted on the system of meaningful differentiation? If so, is the volatility in scores or designations over time reasonable?

2. Claim to be Supported by Evidence: The *state's system of meaningful differentiation* provides accurate, reliable, and valid information about the relative performance of schools in a manner that reflects the state's values and priorities.		
Assumptions	**Research and Evaluation Sources**	**Example Analyses or Data**
• Aggregation procedures are applied accurately. • The intended emphasis of indicators accurately reflects the results of the state's system of meaningful differentiation (e.g., effective vs. nominal weights of indicators) • Overall schools scores/ ratings/classifications are reliable. • Ratings reflect separation in school profiles/data	• Applying accountability business rules to prior year data to compare score/school profiles over time • Quantitative examination of overall index scores (if applicable) or indicator-based profiles • Comparative analyses of indicators to ensure that sufficient variability exists to support.	• Descriptive and inferential analyses of overall index scores or indicator-based profiles (e.g., *k*-means clustering, discriminant analyses, distance comparisons) • Comparing distribution, shape, central tendency, and variability of indicators to other indicators and over time • Quantifying the influence of indicators on overall score/designation variability (e.g., commonality analyses, comparisons of measures of central tendency) • Identify max/min values of overall scores or indicator values and compare to means and standard deviations over time to assess volatility

[4] Meaningful differentiation refers to the degree to which schools' outcomes and characteristics differ based on their designation or category (e.g., A-F rating).

Claim 3: The state's procedures for identification (Comprehensive Support and Improvement, CSI/Targeted Support and Improvement, TSI) are fair and identify schools believed by the state to be most in need of support.

This third claim pertains to the federally required school designations and raises questions regarding procedures, intended outcomes, and rates of identification. While all SEAs were required to propose CSI, TSI, and Additional Targeted Support and Improvement (ATSI) school identification processes, states differ in their TSI and ATSI methodologies. While the considerations in the table below cannot address every state design, the these guiding questions are applicable to the processes and capacity concerns that have emerged throughout the peer review process and in response to U.S. Department of Education feedback:

- Has the SEA developed school policy-level descriptors (SPLDs) for CSI/TSI/ATSI designations?
- Does peer review feedback reflect initial design decisions? If not, have SPLDs been appropriately updated, and do they continue to reflect the state's policy priorities?
- Do identified school profiles reflect underperformance or lack of progress as specified by policy and design decisions?
- Is there clear separation in data profiles for CSI and TSI schools compared to non-CSI and non-TSI schools, respectively?
- Do identification rates of schools fall within the state's capacity to support state-led or LEA-collaboration improvement efforts?

3. Claim to be Supported by Evidence: The state's *procedures for identification (CSI/TSI)* are fair and identify those schools believed by the state to be most in need of support.		
Assumptions	**Research and Evaluation Sources**	**Example Analyses or Data**
• The state has specified School Performance Level Descriptors for CSI, TSI, and ATSI schools. • The state's procedures for identification do not over-identify certain types of schools (N-counts or high free reduced lunch populations). • Schools identified for CSI or TSI demonstrate a profile of performance consistent with that believed by the state to be most in need of support • Procedures used to identify TSI schools capture those schools with subgroups in need of support. • ATSI and TSI identification approaches do not put a strain on state and local capacity to deliver support.	• Stakeholder feedback sessions, leadership vision, SEA priority articulation for CSI/TSI/ATSI schools • Crosswalk of intended identification design and revised identification business rules for CSI/TSI/ATSI schools • Quantitative examination of data over time • Comparative analyses by school category/designation and school profiles to compare indicators and overall performance	• Qualitative feedback from leadership and stakeholders through focus groups or interviews • Crosswalk of support structure design and simulated identification rates; self-assessment of internal capacities for support • Cluster analyses of school types using their indicators as predictors • Simulations applied to historical and projected data to confirm identification of schools and estimate volatility (e.g., bounce across thresholds) • Distance analyses of indicators and measures by school types (e.g., *k*-means or proximity analyses) to show differences among TSI, CSI, and other schools

Claim 4: Exit criteria are attainable and accurately identify schools that are no longer in need of support.

This fourth claim pertains to the federally required exit criteria associated with the CSI, TSI, and ATSI school designations. (Given the district-role for TSI exit criteria, the SEA will need to partner with districts to evaluate expectations statewide and provide support.) These guiding questions should be considered by states as schools are monitored with respect to state-defined requirements:

- Do design decisions for exit criteria reflect the system or Theory of Action's policy priorities (e.g., equity, universal improvement)?
- Have simulations been conducted using historical data? Do data reflect plausible gains over time to exit?
- How do school profiles for CSI or TSI schools change over time compared with changes of non-identified schools?

4. Claim to be Supported by Evidence: *Exit criteria* are attainable and accurately identify schools that are no longer in need of support.		
Assumptions	**Research and Evaluation Sources**	**Example Analyses or Data**
• Exit criteria reflect the outcomes for improvement expected in the system (e.g., overall performance changes, subgroup-specific improvements). • Historical improvements suggest exit criteria are attainable with focused improvement efforts. • As schools exit identification status, their profiles reflect separation from other newly identified schools.	• Alignment between policy decisions, design decisions, and school profiles over time • Quantitative examination of simulation data • Comparative analyses of changes in school profiles (overall, by indicator, by subgroup) between identified and non-identified schools	• Qualitative crosswalk between SEA policy priorities and exit criteria design decisions • Simulations applied to historical and projected data to confirm attainability of exit criteria • Change-over-time comparisons between identified and non-identified schools to determine if acceleration can be expected using historical data

Claim 5: School designations and school data profile information communicate school quality to stakeholders as intended.

This final claim addresses the larger impact of school designations and how school performance is communicated to the public and educators. ESSA requires states to report school performance based on their systems of meaningful differentiation. However, states have a significant amount of flexibility in how they report information. These guiding questions point to considerations regarding coherence between performance standards for schools and how this information is communicated:

- Has the SEA conducted and documented a performance standards setting for school designations? (While this question relates to mechanics associated with the second claim, here it is more focused on the message that school designations communicate.)
- Have reports and school profile descriptions been evaluated by stakeholders?[5]
- What are the hit rates of online reports? What feedback do local sites provide regarding report use?
- Have coherent connections been established between (lagging) outcome data and (leading) indicators? Have these connections been communicated through resources or tools?
- Do reports provide low/no-stakes data in support of local monitoring efforts?
- Does reported school-level information reflect the expectations identified through performance standards setting?

5. Claim to be Supported by Evidence: *School designations and data communicate school quality to stakeholders as intended.*		
Assumptions	**Research and Evaluation Sources**	**Example Analyses or Data**
• Performance standards for the system are defensible. • Overall school ratings support intended interpretations of school performance. • Reports are easily accessible used by targeted groups of end-users • Reports are clear and intuitive • LEAs and schools understand how to interpret and use the results to inform improvement efforts.	• Representative and informed group engages in standard setting using SPLDs, school profiles, and impact data • Policy review includes a "reasonableness check" for performance standards over time • Quantitative examination of web page views and report access • Qualitative feedback for reporting structure and designed reports • Multi-stakeholder group brainstorming sessions to develop resources and tools • Identification of coherent low/no-stakes data by SEA and LEA experts	• Simulated performance standards applied to historical and projected data to determine change over time to impact data • Descriptive analyses of web and report access by location • Focus groups and feedback sessions for reporting tools and structure • Analysis of brainstorming sessions; focus groups review proposed resources and tools • Focus groups and feedback targeting supplemental information provided in reports that is not explicitly linked to accountability data or outcomes

[5] States are encouraged to read the Performance Standards Brief (Domaleski, D'Brot, Keng, Keglovitz, & Neal, 2018) for more detail on setting standards.

Conclusion

This brief presents a high-level conceptualization of how accountability and improvement systems are intended to work together and, further, why SEAs should consider developing a validity argument for their accountability and improvement systems. Additionally, it describes a framework that states can use to evaluate their identification decisions by examining their system-specific claims, assumptions, and potential sources of evidence. These considerations can assist states and their districts in understanding how decisions based on sound theories of action can lead to defensible identification systems that inform—and are informed by—statewide systems of support.

References

AERA, APA, & NCME (2014). Part 1: Validity. *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Council of Chief State School Officers (2011). *Principles and Processes for State leadership on Next-Generation Accountability Systems*. Washington, DC: CCSSO.

Domaleski, C., D'Brot, J., Keng, L., Keglovitz, R., & Neal, A. (2018). *Establishing Performance Standards for School Accountability Systems.* Washington, DC: CCSSO.

D'Brot (2018). *A framework to monitor and evaluate accountability system efforts*. Dover, NH: Center for Assessment.

D'Brot, J., Lyons, S., & Landl, E. (2017). State systems of identification and support under ESSA: Evaluating identification methods and results in an accountability system. Washington, DC: Council of Chief State School Officers.

Goldschmidt, P. (2018). *Handbook for developing and monitoring the English Language Proficiency indicator and English Learner progress*. Washington, DC: Council of Chief State School Officers.

Keng, L., D'Brot, J., & Landl, E. (2018). *Accountability identification is only the beginning: Monitoring and evaluating accountability results and implementation*. Washington, DC: CCSSO.