Assessment Transition and Implications for Accountability

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

Produced for the Council of State School Officers Accountability Systems and Reporting State Collaborative on Assessment and Student Standards

Introduction

Many states are working toward full implementation of the Common Core State Standards (CCSS) with administration of new summative assessments aligned to these standards in the 2014-2015 year. While state leaders have rightly focused on the challenges of development and implementation of new assessments, the impact of these assessments on existing state accountability systems often receives less attention. This is particularly important given that assessment results factor prominently in state accountability determinations.

This paper will examine the implications of shifting to Common Core aligned assessments on state accountability systems. The primary focus is on state-wide school accountability systems, especially those designed to fulfill the federal requirements of No Child Left Behind (NCLB), including systems approved under NCLB flexibility waivers. However, the issues addressed in this paper will likely inform transition planning related to student level accountability and/or educator evaluation systems.

A guiding theme in this document is that advance planning and analysis are critical to a successful accountability transition process. To that end, five central questions that should be included in an accountability transition plan are addressed:

1. **Expected Impact:** What impact are the new assessments expected to have on school accountability outcomes?
2. **Design Decisions**: What inputs should be included in the accountability system and what outcomes result from their inclusion?
3. **Performance standards**: How should the state establish criteria for 'good enough' performance given defined inputs and outputs?
4. **Results**: How should accountability outcomes be reported?
5. **Consequences**: What supports and inventions should correspond to system outcomes?

It is important to acknowledge that state accountability transitions can take many forms. Some states may use the transition period as an opportunity to make substantial changes in the system, guided by the view that a new approach to accountability will better support state policy goals. Other states may seek to minimize disruptions to the system, holding the view that the current accountability approach supports the state's priorities. While the premise of implementing widespread changes to the system is addressed in this paper, the reader should not assume this approach is recommended or applicable in all cases.

More broadly, it is critical to emphasize that any changes to the system should be specified in a transition and evaluation plan and an associated timeline developed in advance. Ideally, the process should include extensive stakeholder input, expert review, and analyses of impact data. Often these activities require a substantial commitment of time and resources, beyond what can be accomplished quickly following administration of the new assessment; therefore, the plan should include the nature and timing of accountability reports and consequences for the first year of the new assessments. This paper concludes with suggested analyses that could be incorporated into such a plan to inform evaluation.

Expected Impact

Given the time, effort, and research states spend developing their accountability systems and communicating the details of those systems to different stakeholder groups, it is understandable why many are reluctant to make design modifications despite a shift to assessments aligned to the common core. Furthermore, upon first glance it may seem as if an existing system will be completely appropriate as is, requiring only that results associated with new summative assessments are swapped in for—or considered in conjunction with—one or more previously administered state tests.

An assessment change can, however, influence the results and inferences associated with an accountability system to a greater extent than one would initially expect. This is due not only to the impact of using new assessment results, but also the influence of broader system-based changes that often occur in conjunction with a shift to new standards and assessments. Consequently, it is useful to acknowledge the likely or known impacts associated with the use of new assessments prior to thinking about how and/or if a state's current accountability system needs to be revised. For illustrative purposes, a few likely implications are outlined below:

1. Assessments associated with given grade/course will be more rigorous.
2. Assessments will require students to display different types of knowledge and skills than previously measured.
3. Students will be held to different, more rigorous expectations in terms of the knowledge, skills and competencies necessary to demonstrate proficiency within a grade/course (i.e., new performance standards and proficiency level descriptors)
4. Student performance relative to defined proficiency standards will decline.
5. Different scales and metrics will be established to evaluate student achievement and calculate growth.
6. The number and type of students included in the general assessment population may change (e.g., due to the removal of modified assessments, improved accommodations, etc…)
7. State-wide programs and initiatives will be put in place to support schools and allow teachers to better prepare for implementation of the common core.

Clearly detailing the implications of shifting to common core assessments makes it easier to think about how and/or why an existing accountability system may need to be revised. For example, if students are being held to more rigorous expectations, the percentage of students

achieving grade-level proficiency standards may be much lower than it has been in the past. Since school accountability systems are heavily influenced by assessment outcomes such as proficiency rates, an expected impact to the accountability system may be that the percentage of schools receiving unfavorable ratings will dramatically increase. Even in the absence of information about performance standards for the new tests, existing external indicators may be useful to roughly predict impact. Such indicators may include the percent of students scoring basic or proficient on NAEP or the percent of students meeting ACT college readiness benchmarks. Similarly, if growth is calculated differently than it has been in the past, standards related to how much growth a school must show within a given year may need to be revisited.

While the list above provides a place to start we strongly recommend that states think about the specific changes that will occur as a result of assessment transition prior to evaluating the appropriateness of their existing accountability system.

Design Decisions – Inputs and Outputs

In order to be useful and defensible a state's accountability system must demonstrate coherence. That is, the design of the system must clearly align with the state's overall goals related to accountability. This alignment should be both transparent and logical, such that all stakeholders clearly understand why specific elements were selected for inclusion. In addition, the design of the system should clearly represent what the state values in terms of achieving desired change (i.e., the state's theory of action related to how the accountability system will bring about change).

For many states, the goals associated with school accountability have changed in conjunction with the shift to the common core. The focus is less on getting all students to "proficient" and more on ensuring students are provided with the skills that they need to be college and career ready upon graduation from high school. The impact associated with this shift has been represented in a variety of different ways across states, including the initiation of interim and formative assessment systems to evaluate student understanding/progress, professional development related to teaching to the common core, and a greater focus on critical thinking.

The inputs to an accountability system are the measures, results or indicators selected by the state to evaluate school performance. The inputs, therefore, represent what the state values and prioritizes as a marker of school progress relative to the attainment of state-defined goals. Test scores and graduation rates are two common inputs in the context of school accountability.

Outputs are the ratings, decisions, classifications or indices that result from the collection and aggregation of input measures. Outputs are used to inform school-level accountability decisions. For example, one or more assessment inputs may be used to provide for output measures related to student achievement, growth and equity (i.e., reducing the achievement gap). Similarly, graduation rate, attendance rate, AP participation rate and other school-level inputs may be combined to provide an output related to overall school health. In most cases, key outputs are

further aggregated to provide for one overarching rating or classification related to school performance.

To evaluate the appropriateness of an existing accountability system, states must review defined inputs and outputs to make sure they are still relevant in light of newly established goals or initiatives. For many states, additional inputs that better reflect the state's goals related to school improvement may be considered. For example, a state that administers performance tasks in mathematics on a quarterly basis to evaluate student performance in mathematical modeling may decide to include these assessments as inputs into their system. Similarly, a state may determine that assessments in science or evaluations of 21st century competencies, such as oral communication, are necessary inputs to support inferences related to school progress in supporting college and career readiness.

Likewise, state outputs identified to support decision making should also be evaluated. States must consider not only what outputs make sense, but also the manner in which associated inputs should be aggregated in light of state priorities and goals. For example, it may be determined that an output related to student achievement should be based 75% on performance associated with the state-defined summative assessment and 25% on performance associated with other measures (e.g., formative assessments, performance tasks, etc…). Similarly, if a state emphasizes growth over status, final school-level classifications may weigh outputs related to growth to a greater degree than they have in the past. The key is being able to clearly show how the selection of inputs and definition of outputs support a clear, coherent message about the state's goals and overarching theory of action related to accountability.

Performance Standards

Performance standards provide a means for differentiating performance relative to a defined set of expectations. Within the context of assessment, performance standards are operationalized in terms of cut-scores and performance level descriptors (PLDs). When a state decides to change their content standards and assessments, best practice dictates that a process be conducted to establish new, or validate existing performance standards. This process typically involves convening committees of subject matter experts to review PLDs in conjunction with test items and student performance data to make cut score recommendations (i.e., define the score-based thresholds that differentiate performance levels). Such activities are necessary to ensure that proficiency expectations are realistic and score-based inferences accurately represent the competencies measured by the test.

In a similar fashion, when the components underlying an accountability system change, the standards which define adequate or expected performance within the context of that system must be reviewed. This is true at the output level (e.g., achievement, growth, etc…), as well as at the overall school-rating level or score.

The shift to assessments aligned to the common core means different things for different states. For some, it represents a significant change in terms of what and how much is expected of students at a given grade. For others it represents a change in how expectations are assessed (e.g., mode of administration, item format, etc…), and/or a different way of thinking about how school progress should be measured (e.g., growth instead of status). Consequently, when defining and/or evaluating performance standards for accountability, states must think about what "good enough" performance means in light of*:* 1) the impact new assessments are expected to have on defined system inputs, and 2) current thinking about how (and to what extent) defined inputs should influence decisions about school performance.

For example, if the assessments and associated standards used to evaluate student achievement are more rigorous than that of previous years, expecting schools to attain previously established proficiency rates may not be reasonable. In this case, lower proficiency rates that account for the expected impact of new content and assessments on student performance in conjunction with a plan to phase in more stringent standards over the next several years may be defined. Similarly, if different sub-groups are expected to respond to the new assessment in different ways, previously defined equity standards may not be appropriate. Consequently, a state may decide not to include outputs related to equity within the system, or to assign them minimal weight relative to a school's overall rating, until the impact of assessment transition is better understood.

Like performance standards for assessments, accountability standards must be established with an understanding of the inferences and implications associated with assignment to a given level. The information a classification is intended to provide defines, in large part, the procedures, data and stakeholders that should inform it. For example, if a school receives an overall classification of B using a school grading metric, what does that mean? Does it mean the school met threshold accountability requirements for 4 out of 5 indicators? Does it mean the school is likely to achieve defined College and Career Readiness (CCR) targets by 2020 if current performance and rates of growth persist? Does it mean that the average weighted performance over defined system outcomes was in the top 30-40% of schools in the state? Or does it mean that the profile of scores obtained by the school represents the type of school-level performance expected by stakeholders and/or policy makers? Clearly, each of these inferences clearly necessitates a different procedure for evaluating and/or setting appropriate standards.

Transition also provides states with an opportunity to consider whether existing performance designations, at the outcome or overall system level, still make sense. Thought may be given to the number of levels used to classify results, their labels, and the manner in which they are defined. For example, a state may decide that it is no longer necessary to differentiate school performance related to student achievement in terms of 5 performance levels, and that 3 will suffice (e.g., Low, Medium and High). Or, if a state wants to make it clear that they have transitioned to an accountability system that is completely different from that which was previously in place, they may use a new system for differentiating overall school performance (e.g., moving from a letter grade system to one based on new performance categories or

changing the scale of a composite accountability score.). This topic is considered in more detail in the "Results" section of this paper.

As with inputs and outputs the key is to consider what makes sense in light of a state's goals related to accountability. If a state plans to use accountability results to identify struggling schools, for example, establishing performance standards that identify all schools as struggling will be counterproductive. Performance standards must be relevant, defensible and reasonable in light of the effect of assessment transition on students, teachers and schools. Furthermore, a plan should be put in place to support ongoing evaluation of the standards as familiarity with the assessment and common core increase.

Results

Another key issue to consider is the extent to which comparability of results is desirable. Specifically, given that accountability results will be comprised of a new set of tests, is it is desirable to produce an end result consistent with that previously provided? Also, should the distributions of outcomes be the same? For example, in a system that reports results as a composite or index score, would it be desirable to maintain an index with the same properties, and should score points or categories on that index have the same meaning?

There are three general approaches to address this issue of comparability: 1) maintain current outcomes and standards 2) maintain current outcomes with new standards, 3) establish new outcomes and standards.

Maintain Current Outcomes and Standards

Just as scales from two different tests can be equated if they share the same construct, it is possible to produce new accountability outcomes (e.g. composite score and/or classification categories) that match the characteristics and standards of the current system. In general, this requires establishing benchmarks for equivalent performance by either defining levels of achievement on the new CCSS tests that are empirically comparable to the standards for the legacy state tests or by computing an overall adjustment to the outcome following production. This could be accomplished by determining the relationship between the new and former school level results produced with each set of assessments via equipercentile concordance, regression or a similar method.

The advantage of this approach is that results will have the same 'look and feel' as that produced by the existing system. Additionally, the performance expectations and resulting distribution of school performance will be generally equivalent. On the other hand, such an approach could be regarded as masking performance deemed below expectations, particularly to the extent that the new tests are more rigorous than the legacy tests.

Maintain Current Outcomes with New Standards

Another alternative is to produce a comparable scale and/or classification system, but establish new standards. This is accomplished by establishing (or validating) a definition of 'good enough' performance with respect to the new tests in accordance with the guidance addressed in the previous section on performance standards. By keeping the design decisions consistent with the state's current approach, outcomes will be reported in a manner that is consistent with the characteristics of the current system in all aspects except for one or more performance standards. A drawback of this approach is that scores that may appear to be comparable will not be meaningfully comparable from the previous system to the new system, which risks misinterpretation.

Establish New Outcomes and New Standards

A third option is to produce a distinctly different overall accountability metric with this new system that incorporates the new performance expectations. The advantage of this approach is that it clearly communicates the existence of new standards and/or model components and stakeholders are less likely to make inferences about year-to-year performance that are not meaningful. Also, this allows states to make sure the system accurately reflects priorities and state goals, which may have shifted since the previous accountability system. On the other hand, this approach creates a new and unfamiliar scale that is inconsistent with the former system, which will likely require guidance and training to support appropriate interpretation and use. However, such initiatives are likely appropriate with any of the above approaches.

Consequences

Ultimately, the purpose of an accountability system is to promote student achievement at all levels. A critical mechanism to help produce this intended outcome is a well-designed system of supports and consequences that are tied to accountability results. These consequences can include rewards, such as recognition or flexibility to waive requirements. In other cases, they may trigger interventions designed to address shortcomings, such as developing an improvement plan or implementing supplemental educational services in targeted areas. Often the assignment of a classification itself can be viewed as an important consequence that may be regarded as positive or unfavorable.

A transition plan should take into account how the consequences will be implemented in the short and long-term. There are two primary questions to address. First, should the existing set of consequences be revised to fit the new model? Second, should the consequences, whether revised or not, be activated and assigned in the same manner?

The first question addresses the straightforward issue of whether it is necessary to modify the types of supports provided to schools in all categories. Because supports are often tied to specific information produced by the system, changes to the system will usually necessitate revisions to the support plan. If the state's expectations for how teachers should be teaching and students should be learning have changed, the support plans tied to the accountability system

should evolve to address these changes. A view that consequences and supports can be 'carried-over' from the legacy system may fall short of promoting the types of outcomes that the system is designed to reinforce.

The second question addresses whether the consequences should be tied to outcomes in the same manner for the new model. If, for example, a state has a letter grade accountability system in which the most substantial consequences are provided to schools earning a 'D' or 'F' and if the number of schools receiving these classifications dramatically increase in the new system, will the same consequences apply? Consider that a D school may have earned a C,B or higher had the legacy system remained in place. The new classification has more to do with the assessment and accountability system than changes in actual student achievement at the school. In fact, schools showing a decline might show improvement under the old system. This begs the question of whether the same system of consequences and supports remains appropriate for the new distribution of school outcomes. Additionally, states should consider if they have the capacity to implement interventions to the degree necessary. Depending on how the state addresses these considerations, it may be appropriate to delay or differently assign consequences during the transitional period and/or for a longer term.

Evaluation

Finally, it is critically important to evaluate the accountability system prior to and following release. An ongoing system of monitoring and evaluation helps ensure that the system is functioning as intended and that outcomes are reliable and valid.

Reliability refers to the consistency or stability of a measure. In this case, states are interested in the reliability of the measures of state, district, school and subgroup outcomes. There are multiple statistical approaches to evaluating the reliability of determinations. However, at a minimum it is advisable to track the consistency of outcomes for various levels (e.g. schools, subgroups) within and across years. Although not without exception, it is expected that results will be well correlated for similar school types within year and for the same schools across years. Dramatic shifts in either classification of schools or characteristics of the distribution will signal a troubling lack of stability that will erode the credibility of the outcomes.

If reliability addresses the extent to which the model provides a consistent answer, validity asks, "Is the answer correct?" Stated another way, to what extent are the results credible and useful for the intended purposes? At a minimum, an investigation of the validity of the model should address the following:

1. Are the results associated with variables not related to effectiveness or generally not under the control of the school?
2. Are the classifications credible?
3. Are negative consequences mitigated?

First, it is important to examine the distribution of scores with respect to variables that should not be strongly associated with outcomes. For example, it would be counterintuitive to find a strong relationship between school size and scores, suggesting that the model does not work well for all schools. Additionally, the relationship among variables in the model should be examined. For example, if academic growth and prior year status (e.g. percent proficient) are strongly related, it suggests that high levels of growth are not obtainable by low performing students. Such findings are implausible and erode credibility of the model.

The second question calls for examination of classifications with respect to external sources of evidence that should be correspondent with quality. For example, the state may have a separate school accreditation system and it would be unexpected to find that schools meeting one set of standards associated with school quality, perform very differently in another system. As another example, one would expect high schools with higher college-going rates to receive more favorable outcomes in a system designed to prioritize readiness for college. Importantly, a validity evaluation will be incomplete without an examination for coherence with other indicators associated with high performance.

Finally, a validity evaluation should address the extent to which unintended negative consequences are mitigated. If potentially troubling outcomes occur such as narrowing the curriculum or declining enrollment in rigorous courses, the validity of the system is threatened. Some of these threats could be examined via survey data or focus groups, while others may be explored by quantitative analyses of extant data. Overall, ongoing initiatives to gauge the extent to which positive outcomes outweigh potential negative side effects will bolster the consequential validity of the system and provide a mechanism to promote continuous improvement.