Using Student Growth Percentiles During The Assessment Transition:

Technical, Practical and Political Implications

Damian Betebenner, Elena Diaz-Bilello, Scott Marion, and Chris Domaleski
National Center for the Improvement of Educational Assessment

October 27, 2014

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, 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.

INTRODUCTION AND PURPOSE

State assessment and accountability leaders share many concerns about the transition from existing state assessment systems to new assessments, prominently those produced by the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced consortia. While these leaders are looking forward to the forthcoming changes in curriculum and instruction, many worry about maintaining accountability systems during the assessment transition. There is no question that there are many common issues and challenges across almost all states, but it is nonetheless critical to tailor advice to the specific contexts under which these accountability systems are operating. States transitioning to PARCC, Smarter Balanced, and other assessments will have unique transition challenges relative to the design of their current state assessment systems. One of the looming challenges is the use of student longitudinal growth information in their school and educator accountability systems, ranging from simple gain score models to more complex value-added and student growth percentile models. Recognizing the importance of context-specific factors in providing technical assistance about states' accountability transitions, the Council of Chief State School Officers (CCSSO) convened separate groups of PARCC and Smarter Balanced states that use Student Growth Percentiles (SGPs) as part of their current accountability systems. SGPs are calculated using regression-based methods for describing students' current achievement (conditioned on prior scores) relative to academic peers beginning at the same starting place (Betebenner, 2009). State leaders wanted to talk with technical advisors and officials from other states wrestling with the transition's affect on existing accountability frameworks. Since the underlying assessments used to compute growth percentiles will be changing in all of these states, this transition will also likely require states to determine whether previous assumptions and inferences about growth taking place at the classroom, school or district level can be supported or sustained. Six states from PARCC and 9 states from Smarter Balanced attended the two CCSSO meetings facilitated by staff members of the National Center for the Improvement of Educational Assessment (NCIEA). The PARCC state assessment and accountability leaders that met in December, 2013 were from Arizona, Colorado, Massachusetts, Mississippi, New Jersey and Rhode Island. The Smarter Balanced state leaders met in July, 2014 and were from Hawaii, Idaho, Indiana[1], Maine, Nevada, Oregon, Washington, West Virginia, and Wyoming. In each of these states, student growth percentiles (SGPs) are incorporated into key leading or primary indicators in school, district and/or educator accountability systems.

Though there are similarities across the different states, the challenges of incorporating growth data across the transition will vary based upon policy requirements, different accountability system designs, field testing requirements, and different approaches used to evaluate growth at the various levels (e.g., classroom, school, or district) in each state. This paper highlights technical, practical, and policy considerations for using growth percentiles during the

1 Indiana is currently not participating in either the PARCC or Smarter Balanced consortia but participated in the meeting with SBAC states as part of their initiatives to better understand growth during their assessment transition.

assessment transition period to support state accountability efforts, and recommends analyses and guidelines to help inform decisions in light of these considerations.

Although the recommendations in this paper are tailored to states that attended the CCSSO-sponsored growth transition meetings, many of these general approaches may be useful to other states whether they are transitioning to PARCC, Smarter Balanced, or something else, and whether they are using SGPs or other growth/value-added models. However, as we noted above, advice presented here should be considered relative to the unique contexts of each state's system.

TRANSITION CONDITIONS

Implementation Timelines

Based on input received from the 15 states that participated in both meetings, many states intend to move forward with using SGPs during transition. However, some states have already moved forward with decisions not to use SGPs for accountability during the transition. For example:

- Colorado intends to run analyses to determine whether the transition SGPs should be used for school and district accountability, but has already informed districts that they are not required to use the transition SGPs for educator evaluations.
- Rhode Island will calculate SGPs for diagnostic purposes but will not report SGPs for accountability since they do not plan on generating SGPs until they have two years of spring-to-spring test results.
- Washington intends to use SGPs for school accountability, but has decided to suspend the use of SGPs for educator or leader evaluations until the 2016-2017 school year.
- Idaho is considering suspending the use of SGPs for accountability for one year during the transition period.

For those states moving forward with generating SGPs for accountability purposes during the first operational year of both consortia (2014-2015), SGPs will need to be calculated using their respective legacy state assessment results as prior scores in the SGP calculation. For those states able to calculate SGPs in 2014-2015, they may also choose to continue using prior scores from their state tests during the second year of operation (2015-2016) because of the added stability associated with including a second prior score in the model.

Figure 1 depicts the timeline associated with the transition to the PARCC assessment. Based upon a thorough review of the timeline at the December Growth Transition meeting, one of the most consequential considerations of this timeline is when scale scores from the 2014-2015 assessment will be returned.

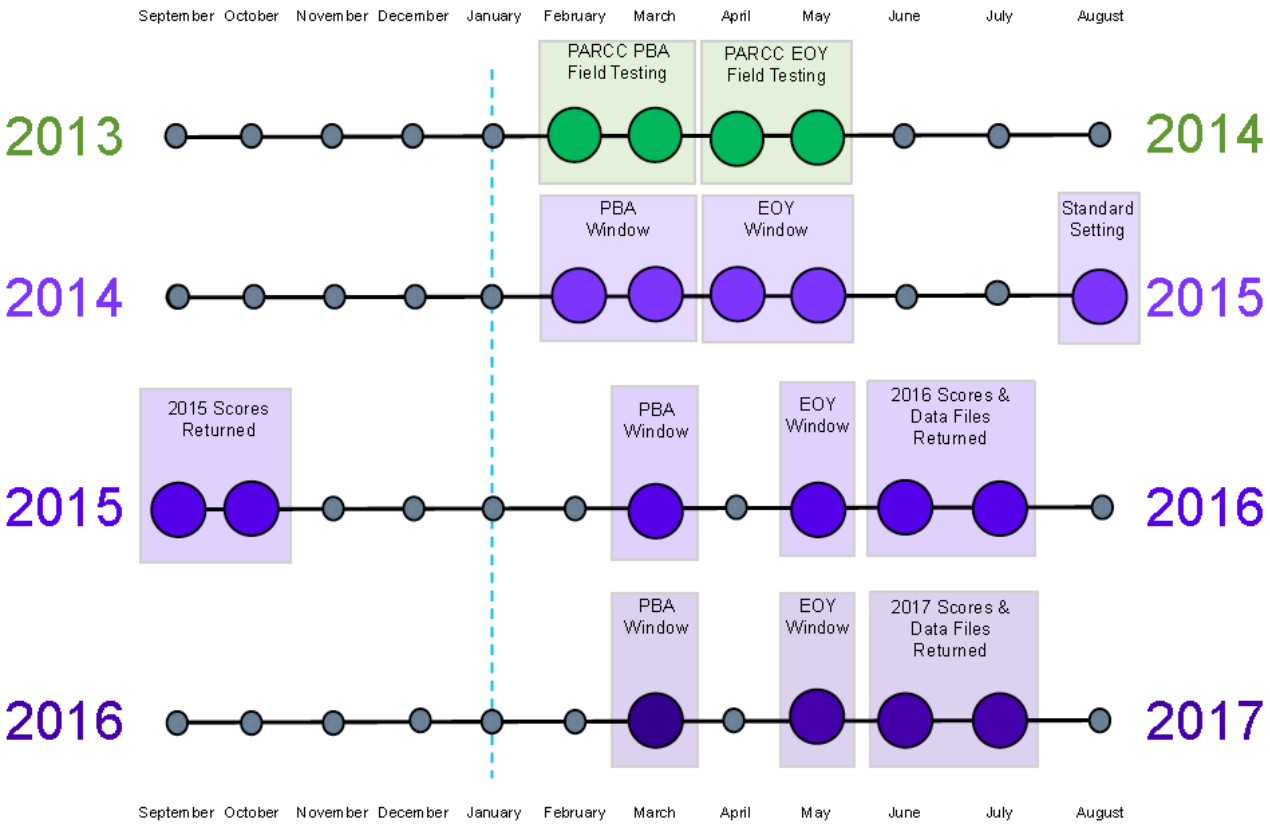

Figure 1. PARCC Transition Timeline

PARCC plans to release consortium-referenced SGPs in the 2015-2016 year and a few of these states may also consider the integration of those results in their accountability frameworks. At a minimum, the PARCC states will need a communication strategy to deal with two different sets of SGP results. All states leaders indicated that they are currently required to return assessment results and derivatives from those results such as growth (either by law or by rule), as early as August 1st. Though the timing associated with the 2014-2015 PARCC rollout is not final, current plans involve returning score reports as soon as possible after standard setting, which would mean sometime in early Fall 2015. However, in response to state requests, PARCC is considering various options for releasing scale scores, without achievement levels, prior to standard setting. At this time, the specifics of the available data and associated timeline await discussion within PARCC.

Figure 2 depicts the timeline associated with the transition to the Smarter Balanced assessment. A major challenge with the transition period identified during the meeting with Smarter Balanced states relates to the production of growth percentiles using 2013-14 results based on the field test designs deployed in several states.

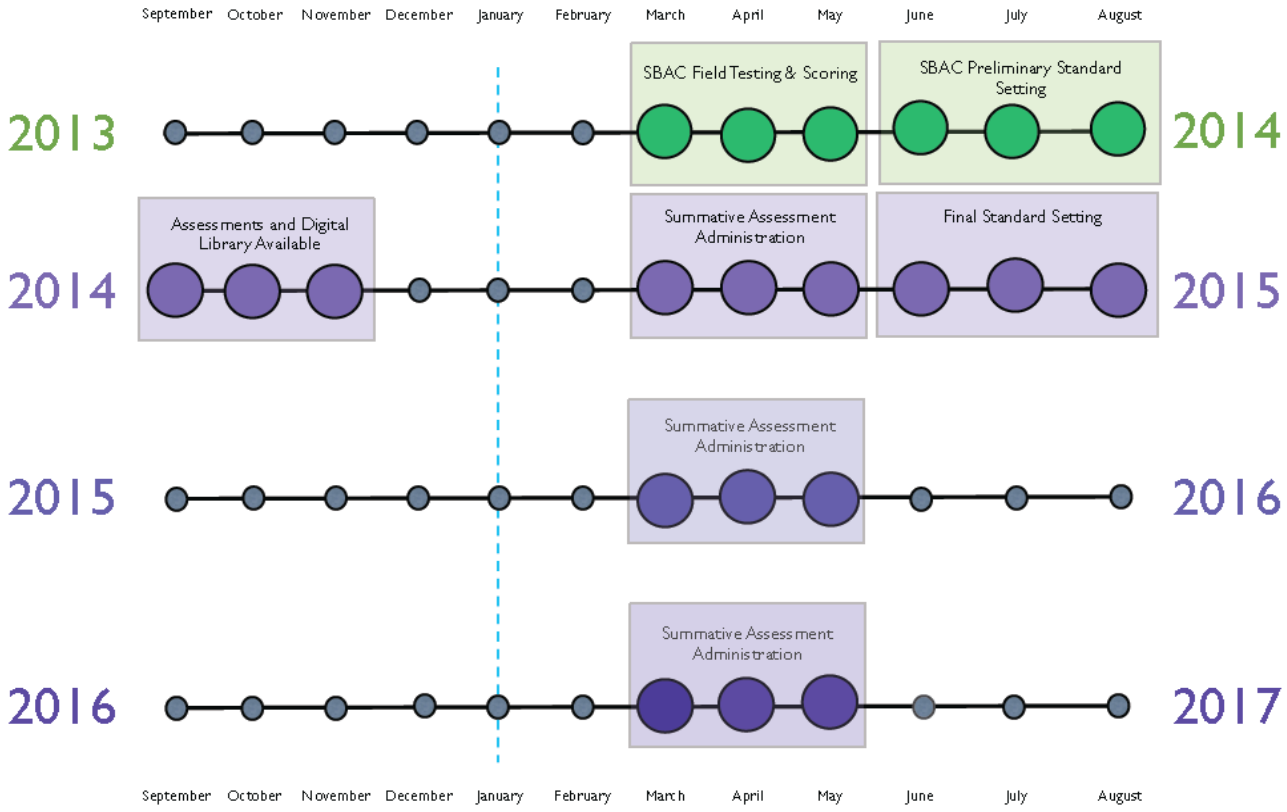

Figure 2. Smarter Balanced Transition Timeline

Calculation of a student growth percentile requires, in general, scale scores from consecutive grades and years for a student. In contrast to PARCC states, some Smarter Balanced states have implemented different field testing strategies leading, in many cases, to missing data for students. Smarter Balanced field testing, like PARCC, returns no student-level data in 2014 with which to run growth analyses. This situation leads to three distinct outcomes occurring in Smarter Balanced states:

1. A state has a **subset** of students participating in the Smarter Balanced field test and these students **do not participate** in the state's legacy assessment.

2. **All** of the state's students participate in the Smarter Balanced field test and students **do not participate** in the state's legacy assessment.

3. A state has a **subset** of students participating in the Smarter Balanced field test and these students also **participated** in the state's legacy assessment.

Unlike status/attainment indicators, missing data impacts the calculation of annual growth of students in both current and future years. Obviously, when the missing data are in the current year it is not possible to calculate an SGP. However, these missing data mean that SGPs cannot

be calculated in the subsequent year since those values from what will be the prior year prevent the calculation of an annual SGP. Due to the considerable challenges of computing growth percentiles with these field test designs, we next weigh the implications associated with each of the three outcomes impacting various Smarter Balanced states.

Implications for Missing Data for Accountability in Smarter Balanced States

Outcome 1: A state has a **subset** of students participating in the Smarter Balanced field test and these students **do not participate** in the state's legacy assessment.

This outcome, associated with the majority of Smarter Balanced states, leads to missing SGPs in both the current and coming year. In some states, with field testing percentages in excess of 25% of the students, this represents a large number of missing data points. In those states using growth for accountability purposes, these missing data can impact the accountability system in multiple ways including:

1. If the sample of students taking the Smarter Balanced field test from the state is not "random", then the growth calculated for the remaining students participating in state testing will be systematically different, making for different comparisons from previous years. States should consider whether growth results should be re-normed or anchored to previous year's norms such that comparisons of SGPs or aggregate SGPs to previous years should be avoided for both 2015 and 2016.

2. Going forward, two different cohorts of students will exist with different testing histories. At some point, the state will need to decide when to "sunset" the use of prior results from the state assessment, and run analyses solely based upon Smarter Balanced results. At a minimum, this will occur when the state has two years of operational Smarter Balanced results available.

3. Districts, schools, instructors, etc. whose evaluation depends, at least in part, on student growth will not have that component available for the students with missing data. In cases of educator evaluations, some educators who previously received evaluations including the growth of students will not have that growth component available.

Outcome 2: **All** of the state's students participate in the Smarter Balanced field test and **do not participate** in the state's legacy assessment.

This outcome where universal field testing occurs with no double testing, was indicated by only one attending state (Idaho), but may be a design in place for other Smarter Balanced states (e.g., CA). This outcome prevents any annual SGPs from being calculated, but allows for bi-annual growth for all students with no bifurcation of the state population based upon them taking different testing tracks (see Option 1). The substitution of bi-annual growth for annual growth in accountability systems should be carefully considered. Impact studies using currently

existing data can be done to help understand the impact of such a substitution. For example, states considering using SGPs that span more than a single year in lieu of annual SGPs can utilize currently existing data to calculate SGPs for that span (ignoring the annual structure of their current data) and investigate the relationship between these multi-year SGPs and their annual counterparts.

Outcome 3: A state has a **subset** of students participating in the Smarter Balanced field test and these students also **participated** in the state's legacy assessment.

This outcome is associated with a small minority of Smarter Balanced states but is most consistent with PARCC states. It allows for annual growth to be calculated using the state assessment as the prior scores and the Smarter Balanced assessment as the current score in the transition year. This option is the easiest option in terms of continuing an accountability system with growth uninterrupted in the transition year, as all the SGPs will be calculable barring any issues associated with the testing itself (e.g., floor and ceiling effects).

In general, the transition to Smarter Balanced and the existence of missing data in some circumstances will lead to missing SGPs, which can impact accountability systems that rely upon that data. Understanding how widespread such missing data is, whether such data are missing at random or systematically missing, whether bi-annual SGPs will be used in place of annual SGPs, how the missing data will impact the calculation of SGPs going forward, and how the data will be used based upon these considerations are some of the critical issues Smarter Balanced states need to consider.

Summary of Growth Percentile Use for Accountability for Attending States

Prior to the meeting with PARCC and Smarter Balanced states, we collected information from each state to better understand the policy and implementation context in each locale[2]. Table 1 summarizes information collected from each state from the two consortia in three areas: 1) current growth approaches used to support accountability decisions; 2) the policy and contractual areas likely to be impacted by changes in how growth will be used to inform accountability decisions; and, 3), the specific accountability systems in which growth is being used.

2 Table 1 does not include data from one participating state (West Virginia).

	Growth Approaches Used in Accountability			Policy or Contractual Areas where Growth Results will be Impacted by the Transition				Growth Results Used in Accountability Systems Impacted by Transition				
PARCC States	Normative	Criterion-referenced	Baseline-referenced	ESEA Waivers	State Legislation or State Board	School Improvement Planning	Charter Contracts	Educator Evaluations	Leader Evaluations	School Accountability	District Accountability	School Improvement
Arizona	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Colorado	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Massachusetts	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓
Mississippi	✓	✓		✓	✓			✓	✓	✓	✓	
New Jersey	✓			✓				✓	✓	✓	✓	
Rhode Island	✓			✓			✓	✓	✓	✓	✓	
SBAC States												
Hawaii	✓			✓		✓	✓	✓	✓	✓	✓	
Idaho	✓			✓	✓	✓	✓	✓	✓	✓	✓	
Indiana	✓	✓		✓	✓			✓	✓	✓	✓	✓
Maine	✓			✓						✓		✓
Nevada	✓	✓		✓		✓		✓		✓		
Oregon	✓	✓		✓		✓				✓	✓	
Washington	✓		✓	✓		✓				✓		
Wyoming	✓	✓		✓	✓	✓				✓		

Table 1. Impact of Transition in States for Accountability

As indicated in Table 1, there are three primary approaches used by states to evaluate and report growth using the SGP framework in their accountability systems: norm-referenced growth (student growth percentiles), criterion-referenced growth (percentile growth trajectories) and baseline-referenced growth (baseline/anchored student growth percentiles) (Betebenner, 2009). We describe next how each of these approaches is typically used by states in the context of accountability.

Norm-referenced Growth refers to the common approach used by many states to summarize growth achieved at the school, classroom or district levels using the median or mean of the individual student growth percentiles calculated for each grade-by-content area norm group in the state. This approach was first developed in Colorado and subsequently extended to include criterion-referenced growth. Accountability determinations are then made by summarizing the individual level SGPs using the median or mean and classifying the "average" growth achieved at each unit of analysis into various rating categories (e.g., low, typical, or high "average" growth). The number of classification categories set in accountability frameworks and the cut-points used to differentiate levels of growth performance vary across states.

Criterion-referenced growth is used to characterize the norm-referenced growth information provided by student growth percentiles relative to a criterion-based target defined most often by a proficiency standard of interest[3]. At the individual level, the result is a set of student growth percentile targets (target SGPs) indicating the amount of annual growth necessary for

3 This proficiency standard could be set at various standards such as: proficiency, advanced, or partially proficient.

a student to reach or maintain a specified achievement outcome in a specified amount of time (e.g., proficiency within 3 years). This growth-to-standard implementation is often used by states to determine whether students are making growth sufficient to catch up/keep up to proficient status or move up/stay up to advanced status. Some states are using these data to calibrate their systems toward the goal of system-wide growth capable of getting students to career and college readiness as well as monitor movements made over time in the percentage of students falling into the "catch up" and "keep up" categories" for school improvement planning and school accountability purposes.

Within the context of school and district accountability in several states, the median growth percentile (MGP) target needed for students at a school or district to achieve proficiency within three years is evaluated relative to the norm-referenced MGP achieved by each school or district. If the MGP achieved for a given school or district is below the MGP target needed to achieve proficiency, the school or district's growth is assessed using a higher bar or expectations for growth relative to other places where growth achieved either meets or exceeds the MGP target. A few Smarter Balanced states (e.g., Oregon and Idaho) are using this approach based on the accountability model originally conceived in Colorado, where one of two rubrics is used to rate a school or district's growth depending upon whether the school or district met the MGP target. In the rubric used for schools and districts where the MGP is equal to or higher than the target, the performance expectations have lower growth cut-points set relative to the rubric used for schools and districts where the MGP falls below the target.

Baseline-referenced growth is an approach used to evaluate growth by a student, school, district, or a classroom relative to a fixed "baseline" or "anchor" set of years established by the state. Growth percentiles for the majority of states are re-normed on an annual basis. Baseline-referenced data are used to enable the comparison of growth results over time and to permit the state to detect whether statewide growth is increasing from year to year. States using this approach anchor the norm groups each year to a baseline ensuring that the SGPs produced are interpreted relative to the baseline year.

Based on the information summarized in Table 1, all of these states face a set of challenges for using growth percentiles to support accountability inferences and decisions during their transition to the new assessments, whether they are using norm, baseline, and/or criterion-referenced growth:

1. **Any accountability design changes implemented as a result of the transition may seriously affect multiple stakeholders at the state, district, and school levels.**

Any potential changes to the design of accountability systems using growth data has clear implications for policy at both the state and district levels, and presents challenges for communicating those changes to all relevant users. There was universal agreement at the meetings that the foremost challenge was in communicating the changes to stakeholders. All

states will need to communicate how growth will be used during the transition period and when certain approaches can be reinstated (e.g., criterion referenced growth or baseline referenced growth) at a future date. Furthermore, the transition period may have legal implications since the use of growth in the accountability reporting system prior to the transition will need to be repurposed or reframed. For some states, this would require approval at the legislative and/or board levels. In some cases, this would also require re-negotiating performance objectives specified in charter school contracts[4].

2. **During the transition period, all 15 states will need to consider how to interpret the growth achieved by students using norm- and criterion-referenced approaches.**

All states will receive student-level growth data from their respective consortium from the 2014-2015 school year allowing each state to generate SGPs using the new data in combination with their old state assessment data as priors. Many states have already used prior and current scores from different tests to generate SGPs — for example, in situations when standards and assessments have been modified. However, the interpretation of growth results relative to the difficulty and complexity of the new content being assessed by the Common Core State Standards (CCSS) will need to be evaluated and considered by states opting to use those results to support accountability inferences. For the three PARCC states using or considering criterion-referenced approaches, none of these states are planning to move forward with this approach during the transition period. However, the five Smarter Balanced states that are interested in maintaining the use of criterion-referenced growth approaches in their accountability systems, will be faced with the additional challenge of having to determine whether and when their criterion-based targets are defensible in light of changes in test content and standards for proficiency that will likely become more rigorous. Finally, the baseline-referenced norm group approach used in Massachusetts and Washington will need to be suspended until the baseline years can be reset using data from the same assessment system. This will likely require waiting at least until the 2016-2017 to report baseline-referenced SGPs.

3. **State will need to find time and resources to dedicate to carefully examine impact data in order to build defensible accountability systems.**

Evaluating impact data to help support decisions on how to best use and report growth during the transition period is critical and will require both resources and time for states to consider accountability reporting options. Such analyses would include comparing any SGP and aggregate SGP calculations from 2015 with previous year SGP results to gain on understanding of how different types of students and school perform in 2015 compared to

4 For more information about the broader issues associated with assessment transition and implications to accountability see: Domaleski, C. & Hall, E. (2014) Assessment Transition and Implications for Accountability. A white paper developed for the CCSO Accountability Systems and Reporting State Collaborative on Assessment and Student Standards. Washington, DC: CCSSO. Available at: http://www.nciea.org/publications-2/

prior years. States will vary in terms of the accountability design options they may pursue depending on what the impact data conveys when reviewing growth results generated prior, during and after the transition period, and likely after at least three years of PARCC or Smarter Balanced data are available.

As states receive the first year of PARCC and Smarter Balanced data, all states will need to conduct analyses to ensure that growth can be calculated from their prior tests to the current PARCC or Smarter Balanced tests. In particular, a specific threat to the calculation of SGPs is the existence of prominent floor effects on these assessments. If a state's data yields a sizable percentage (> 5%) of students in any grade/content area receiving the lowest obtainable scale score, calculation of uniformly distributed percentiles will likely be difficult, if not impossible. Depending upon the extent of floor effects, the issue might be state specific or may apply consortium wide. Moreover, as the initial year's transition results become priors in subsequent years, issues associated with SGP calculation due to floor effects could continue as long as such data is used in the calculations. This would be true of essentially any regression-based approach for evaluating the change in students' scores such as value-added modeling. This is an example of where accountability methods cannot "fix" certain serious issues with assessment results.

Additionally, if the correlation of state-to-consortium assessment results is lower than expected (e.g., 0.6), the intended inferences about student growth from the legacy assessment to a consortium assessment may be suspect. One could argue that such less than perfect correlations may help identify schools or districts that have successfully implemented the CCSS compared with those that have not. While this may be true, it requires more than SGP results to support such an inference. Further, to the extent that CCSS implementation is associated with district wealth, incorporating transition growth results into accountability systems may threaten the credibility of the fairness of the accountability determinations because stakeholders might recognize that the growth results are related to the resources available to districts and not necessarily to educator or school effectiveness. On the other hand, it can be argued that the growth results may validly illuminate important differences in the fidelity of CCSS implementation.

States should conduct many of the types of analyses described in this document to better understand these and other potential threats. Further, the presumably more rigorous standards on the Smarter Balanced or PARCC assessment will likely lead to unique circumstances that will make the assessment transition period particularly challenging when using growth results for accountability purposes. Unique local contexts and policy choices also complicate the decision to move forward with using or suspending the use of growth percentiles in a given state. For example, in the case of Massachusetts, all districts during the 2014-2015 school year have the choice to continue using the state assessment or administer the PARCC assessments. This policy will likely complicate efforts to compare and make inferences about growth achieved between districts, because the state will have to calculate two different sets of SGPs, each with a median of 50. That is, if the growth results for districts opting to use the PARCC assessments

during the transition period are higher on average than the growth results for districts using the state assessments, comparative policy statements made about which schools exhibit higher growth performance cannot be supported since the underlying measures differ between those schools. Although these challenges apply to the 2014-2015 year, these challenges will not dissipate in the 2015-2016 year. Assuming all students take the same assessment, they will have two different sets of possible priors. This assessment choice policy also brings up the question of whether the same cut-points set in the current school and district accountability frameworks should apply to districts using different assessments. Massachusetts educational leaders will have to carefully review the impact data and other analyses from the two sets of districts to consider adjustments to accountability policies for the 2014-2015 and 2015-2016 school years.

In the case of other states such as Nevada, West Virginia, and Oregon, these states would prefer to maintain the use of criterion-referenced growth during the transition period, in particular to continue monitoring the extent to which disadvantaged groups are making adequate growth toward reaching proficiency targets. Although there are a few approaches (described in the next section) that may allow for these states to continue using the criterion-referenced approach, the inferences made about student growth using any one of these options carry a different meaning than the inferences supported using student growth based on the legacy assessments. Even for places such as Colorado, where the decision was made to remove criterion-referenced growth during the transition period, removing this criterion piece from growth has brought up some concerns about tracking progress made by sub-groups since evaluating their growth relative to the proficiency standard serves as an important indicator for tracking equity for these groups.

As indicated throughout this section, impact data will be especially critical for states to evaluate in order to support accountability design choices during and after the transition period. The next section outlines recommended analyses for using growth to support accountability decisions, and for potentially redesigning accountability frameworks prior to the transition.

PERIOD

A common concern expressed by the 15 state participants regarding the assessment transition is whether it is even possible to calculate growth using two different tests. The simple answer is that it depends on the type of growth one wants to calculate. The change in assessments impacts the calculation of growth in different ways depending upon the type of growth to be calculated and the type of scale present in the state. However, as indicated earlier, considering the different approaches being used by each state to compute growth (refer to Table 1), the change in assessments may warrant consideration of different design options to adjust current accountability systems.

Overall, the technical issues associated with the calculation of growth are *not* as formidable as they have been previously with some states using SGPs. This is due to the fact that many of the technical challenges encountered have been addressed through the inclusion of new functions built into the SGP analytic software (Betebenner, Van Iwaarden, Domingue, and Shang, 2014). A much bigger challenge is communicating results during the transition and deciding whether or how to implement different options for redesigning the current accountability frameworks. A key design consideration in the development of the Student Growth Percentile Model was to build a durable metric that could be maintained across assessment transitions. The extent to which substantial or minor adjustments must be made to accountability frameworks during the transition period again depends on the type of growth analyses used to support each purpose. In the sub-sections that follow, we discuss accountability framework dependencies relative to the growth approach used by each state.

Student Growth Percentiles

Despite the absence of a developmental (i.e., vertical) scale in the transition year with which to measure growth magnitudes, there is nothing to prevent the calculation of the norm-referenced SGPs. Norm-referenced SGPs measure student progress relative to "academic peers" who have the same achievement history as other students in the state. This is unlikely to be a concern since all students in each of the 15 states, except perhaps in Massachusetts, will continue to take the consortium assessment instead of their current state assessment.

Similar assessment transitions have occurred in numerous states in recent years. In 2010, Arizona changed its vertically-scaled math assessment to a new vertically-scaled assessment. After some preliminary analyses showed that neither assessment had significant ceiling/floor effects that could potentially undermine the calculation of SGPs, the state proceeded with calculating SGPs on the new system. Colorado recently transitioned from using the Colorado English Language Assessment (CELA) to the new Accessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS) assessment from the World Class Instructional Design and Assessment consortium. In order to justify the continued generation of growth percentiles for ELL assessments, the state ran analyses to check for ceiling and floor effects and to

identify the extent to which prior CELA results were correlated with ACCESS results. The state also enlisted content specialists to evaluate and compare the similarities and differences in constructs assessed by both tests to justify inferences about growth in language development. Colorado has also been calculating growth from its state assessment to the ACT assessment, which has been administered universally in the state in grade 11 for a number of years in an effort to gauge growth toward college readiness. In essence, if the analyses and the review of constructs assessed supports the use of a different test as a prior, accountability systems developed to evaluate growth using a strictly norm-referenced approach do not require much, if any, additional work to modify the existing accountability framework. If the results of the analyses do not support the use of a different test as prior, the state may need to delay using SGPs in accountability until the 2015-2016 school year after two years of consortium assessment data are available.

When new assessments are introduced, the norm-referenced properties of SGPs will be preserved, insofar as the distribution of SGPs will be uniform for all students statewide. However, it should not be assumed that growth estimates are 'interchangeable' with the growth outcomes students or schools would have received if the legacy state assessment were to continue. While the statewide distributions will be similar, the specific growth estimates that students and schools receive may well differ. This should be expected in the same way that one would not assume performance level classifications will be unchanged when a new assessment is introduced. For example, it stands to reason that a school with relatively strong performance on the new assessments will receive a more favorable growth estimate in the year following the transition compared to prior years, just as they would in any year with strong performance.

An accountability design consideration that should be investigated during the transition period with norm-referenced SGPs is the reporting of pooled SGP data across years. Pooling the SGP data entails reporting an MGP based on individual SGP results aggregated across two or more years. Colorado currently uses this approach for their accountability frameworks. The strategy of pooling data has the added advantage of stabilizing the year-to-year growth results reported for accountability purposes. It may also help guard against potentially large performance fluctuations during the transition. However, one area to consider when pooling data is that this approach will mask any year-to-year performance changes since the average performance achieved across years is reported.

Percentile Growth Trajectories

Percentile growth trajectories (also referred to as student growth projections) are a criterion-referenced growth metric on the same percentile scale as the SGP. They allow stakeholders to interpret the norm-referenced SGP in the criterion-referenced context of state assessments and determine whether the student's growth is sufficient to put them on track to reach/maintain a desired level of achievement. Unlike student growth percentiles, which are retrospective and quantify what has occurred, percentile growth trajectories are forward looking and quantify what needs to occur in order to reach/maintain a desirable achievement outcome.

Percentile growth trajectories utilize the coefficient matrices calculated as part of the student growth percentile analyses to project forward along all 99 potential percentile trajectories for a student to determine what level of growth is required to reach various specified achievement outcomes (usually cuts between achievement levels) in specified amounts of time (annually going forward). As such, projections forward along a new test scale require coefficient matrices calculated relative to that scale. In the 2014-2015 transition year, dependent variable scores will be on the new (e.g., PARCC or Smarter Balanced) scale and independent variables will be on each state's previous assessment scale.

As indicated earlier, several states in attendance expressed interest in maintaining the use of student growth projections or criterion-based growth as part of their accountability designs. These states will need to reconsider how growth can be used for accountability during the transition. As several states (in both PARCC and Smarter Balanced) rely upon percentile growth trajectories/student growth projections/student growth targets as part of their current accountability systems, one option for these states to consider is to statistically adjust the new test scale to create the best estimate of the prior test for purposes of continuing to produce comparable growth-to-standard outcomes. The downside of this approach is that the state may desire to model growth to a new and likely more rigorous standard and this approach simply delays incorporating the new expectation. That is, growth targets pegged to the performance standard on the legacy test, may not represent the academic outcome policy makers want to prioritize. On the other hand, carrying the legacy standard forward will minimize 'disruption' in the model, which may be desirable if even for a limited time. In other words, this would essentially translate the consortium test results back to the legacy scale so that it could be seen as saying "our students really performed well" and could delay the discussion of the new expectation. Conversely, one could adjust the old test scale to create the best estimate of the new test to produce new growth-to-standard outcomes. These would not be comparable to the legacy growth-to-standard estimates, but would facilitate continued inclusion of this component in the accountability model. Both of these approaches will have communication challenges, but states that wish to incorporate the new standards in the model as soon as possible will likely find one of these approaches appealing.

A number of approaches for producing statistically comparable scores include the use of an equipercentile concordance approach between the state's assessment scale and the new PARCC or Smarter Balanced scale[5]. Additionally, a number of equating approaches might be feasible if it is possible to have items in common between the tests or have a representative group of students take both tests. Using these equating approaches may provide a temporary solution for ensuring that the coefficient matrices used to project forward are based on an equivalent scale. This temporary solution may be applied in 2015-16, when more than one year will be available for a state on the new assessment scale, before the projections could revert

5 Note for states using the current SGP package: There are plans for the SGP package to be augmented to allow for interim projections to be calculated using an equipercentile concordance between the states assessment scale and the new scale.

back to using assessments from the same system. There are many technical reasons why we would not recommend this approach as a permanent solution to employ, but these are beyond the scope of this paper.

Another class of approaches involves identifying a suitable substitute for the growth target that does not require comparable test scales. To be clear, this alternative would break longitudinal comparability between the old method and the new method, but it may be an appealing option insofar as it allows the state to maintain criterion-referenced growth in the accountability system without statistically adjusting the old or new test scale. Perhaps the most straightforward way to accomplish this is to identify a new growth standard, such as a school MGP target. This target could be identified through a combination of policy and data analyses to provide evidence that the target is sufficient for the purpose for which it is designed. For example, the policy objective may be to select a target based on the legacy assessment such that the majority of students who are below proficient who grow at this rate become proficient in one year or some other time period. Given this definition, it would be straightforward to produce data that indicates the percent of non-proficient students in the prior year classified as proficient in the current year for each MGP.

Baseline Student Growth Percentiles

Massachusetts (as well as Georgia and Washington) makes extensive use of baseline-referenced student growth percentiles in their accountability systems. Student growth percentiles are a norm-referenced growth metric that are almost always normed using the most recent year's data. As such, in a given year the median SGP for each content area/grade combination is 50. Some argue that re-norming growth every year in an accountability system results in a "zero sum game." To rectify this situation, the Massachusetts Department of Elementary and Secondary Education established baseline growth norms against which future cohorts of students' growth would be calculated. Based on the strong assumption of equivalence of grade and content area scales from year-to-year, the growth results can be "anchored" or "baseline" referenced such that the norm for the state is no longer fixed at 50 and that growth is interpreted relative to the specific baseline year of interest. Using this methodology, Massachusetts has been able to demonstrate higher rates of growth over time that are consistent with increasing the efficacy of the education system.

Considering the strong scaling assumptions employed to justify the use of baseline-referenced SGPs, we would not recommend using this approach in accountability frameworks until at least four years (two years to establish the baseline and two years to calculate growth against this baseline) of data are available from PARCC or Smarter Balanced to investigate how well these assumptions hold. In the specific case of Massachusetts, the baseline-referenced approach would need to be suspended for accountability purposes until more data from PARCC can be accumulated.

RECOMMENDATIONS

To summarize key recommendations as well as to highlight important analyses associated with the calculation of SGPs going forward, we recommend that states investigate performance results during the transition years in the following ways:

1. For Smarter Balanced states with field test designs related to Outcomes 1 and 2 (see pp. 6-7), the extent of the missing data will vary by state based upon their field testing plan. It is recommended that in states where missing SGPs will occur and those missing values will impact accountability systems, an inventory of the impact be conducted across the next several years to understand what will be available and how the state will accommodate the missing data across those years.

2. Examine the correlations between the pre- and post-scores used for students (e.g., a grade 5 math legacy assessment and a grade 6 PARCC or Smarter Balanced math assessment) and flag any correlations below 0.6 for inspection. Student growth percentiles can be calculated when scores are uncorrelated, but in such a situation, the conditional distribution associated with a fixed score is equivalent to the unconditional distribution, so the prior score supplies no information. In other words, when scores are uncorrelated across years, SGPs would provide no information beyond status scores. When correlations fall below 0.6, it may indicate that there has been construct shift that should be recognized as part of SGP analyses.

3. For states that do not currently pool SGP data, explore the use of pooling across two or more recent years when reporting the norm-referenced results for accountability. As indicated earlier, this approach may mitigate the impact of performance fluctuations occurring during the transition period and may improve the precision of average growth estimates being reported for accountability purposes. This recommendation might be explored in general, even beyond the transition, to improve the stability of the outcomes.

4. Explore the use of results associated with replacing the growth target or developing comparable scales to evaluate whether this approach may support the continued use of student growth projections or criterion-based growth approaches in the accountability system during the assessment transition. As noted earlier, this should be considered a temporary patch until at least one additional year of data is available to support the projections using scores from just the PARCC or Smarter Balanced assessments. Examining the results to determine whether certain assumptions hold is critical, and although the specific analyses for evaluating these results are not addressed in this paper, we would strongly encourage states to ensure that they conduct these analyses first to determine whether the criterion-based approach can be maintained without interruption during the transition.

   Additionally, since it would be difficult to make the claim that these different assessments are written to the same content specifications, the "equated scores" reported need to be

understood as establishing concordance between test scores rather than equating scores across assessments. Considering that growth inferences based on reported concordant scores do not share the same interpretation as growth inferences made using the legacy assessments, the issue of whether these equated transition scores should be reported for use in accountability systems will need to be discussed with stakeholders.

5. High school math course tests will be employed with the new PARCC assessments and this will require those states to investigate and find common course patterns across grades in order to generate SGPs. Since high school students can take different math course tests at different grades, states will need to evaluate the goodness-of-fit of the data relative to each of the common course patterns identified and ensure that adequate sample sizes are available to justify the selection of a specific pathway. Of course, this tends to be less of a concern for mathematics compared with subject areas such as social studies and science, because math course-taking tends to follow a predictable sequence, except for opportunities like honors and related courses.

6. Lastly, states should investigate the distribution of scores on the new assessment scale by grade and content area, paying particular attention to percentages of students scoring at the lowest obtainable scale score (LOSS) of the scale. The non-parametric, quantile regression-based methodology underlying the calculation of SGPs and percentile growth trajectories is robust to most issues associated with assessment transition and the changing of scales. In other words, the SGP model is essentially agnostic to most linear scale transformations (Betebenner, 2009) so that the new scaling, with possibly different interval properties than the legacy scale, is unlikely to affect calculation. However, if a large percentage (greater than 5%) of a cohort scores at the LOSS of the test, it becomes difficult to calculate SGPs for students that conform to a uniform distribution as is theoretically expected with percentiles. The non-parametric b-spline methodology associated with the SGP methodology allows the model to fit the data and recognize the LOSS and the highest obtainable scale score (HOSS) of the scale. Although corrections for LOSS/HOSS issues have been implemented as optional adjustments within the SGP package, if a preponderance of observations occurs at the LOSS for the new assessments, it may still be difficult to distinguish between student achievement outcomes leading to student growth percentiles when the data do not follow a uniform distribution. Given the design requirements for Smarter Balanced and PARCC, we do not anticipate LOSS to be a problem, but we are flagging this as a potential problem for many other states that are developing their own CCSS assessments. If significant percentages of students score at these levels, SGPs should be examined for model bias to ensure that disproportionate percentages of SGPs are not being assigned to students based upon floor effects of the test.

SUMMARY

All states participating in the December 2013 and July 2014 Growth Percentiles During the Assessment Transition meetings intend to calculate growth annually during the transition to the new consortium assessment. Calculating growth during the assessment transition to PARCC or Smarter Balanced will require careful technical oversight to ensure that results from the SGP analyses can continue to support accountability determinations, and that the design of the accountability frameworks is modified based on results found evaluating the new assessment data. In light of the recommendations provided in this paper, it may also be appropriate to suspend or change the consequences of the model, particularly if a "transitional" phase is in place before the model is regarded as complete. That is, a full or partial hold-harmless phase may be advisable during 'rebuilding' until such time as the model is fully specified and validated.

Although the set of analyses recommended in this paper should not be complicated for each state to perform, the bigger challenge for all states will be to develop a thoughtful communication strategy to help stakeholders understand anticipated changes made to growth, if any, to their existing accountability systems. As noted by Domaleski and Hall (2013), "advance planning and analysis are critical to a successful accountability transition process" (pg. 1). If states do not take the time to engage in this advance planning work, they will likely face considerable difficulty having to rebuild stakeholder trust and ownership in the use of growth as a key feature to each state's accountability system, regardless of whether any changes have or have not been implemented as a result of the transition.

REFERENCES

Betebenner, D. W. , Van Iwaarden, A., Domingue, B, Shang, Y. (2014). SGP: Student growth percentile and percentile growth projection/trajectory functions. (R package version 1.2-0.0) http://cran.r-project.org/web/packages/SGP/.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. Educational Measurement: Issues and Practice, 28(4):42-51.

Domaleski, C. and Hall, E. (2013). Assessment Transition and Implications for Accountability. Retrieved February 20, 2014, from www.nciea.org.

Council of Chief State School Officers

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