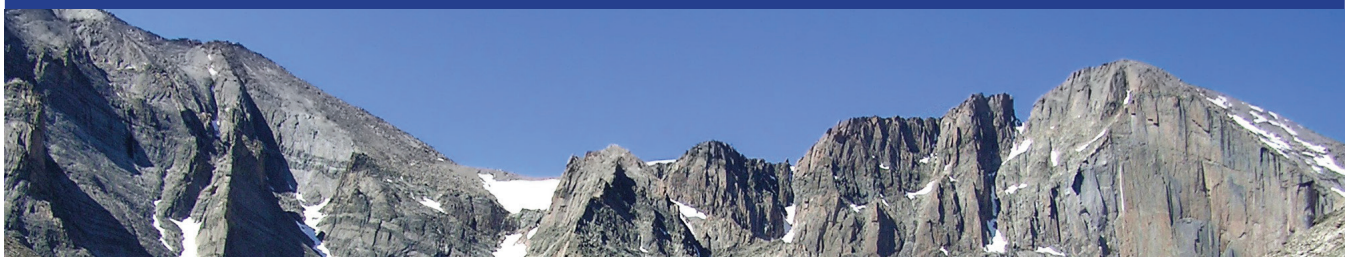
Using Student Growth Percentiles for Educator Evaluations at the Teacher Level:

Key Issues and Technical Considerations for School Districts in Colorado

Elena K. Diaz-Bilello & Derek C. Briggs

JULY 2014



Center for Assessment

CADRE

Acknowledgements

We thank the staff at the Colorado Department of Education and Damian Betebenner (Center for Assessment) for their review of this brief and feedback. We extend a special thanks to Britt Wilkenfeld at the Colorado Department of Education for her valuable editorial contributions.

We also thank Karen Herbert at the Denver Public Schools who authored the DPS eligibility rules located in Attachment A, and Margaret Ruckstahl who provided the Harrison School District Two eligibility rules located in Attachment B.

Introduction

A common feature of educator evaluation approaches developed over the last five years in states and districts throughout the United States is the combination of teacher practice and student performance measures. In Colorado, Student Growth Percentiles (SGPs) generated by the Colorado Growth Model for mathematics, reading and writing are a key component that can be incorporated into districts' educator evaluation systems. Currently, districts in Colorado may be considering approaches for SGPs for use as part of an educator's evaluation. The mean or median of Student Growth Percentiles, an MGP, is just one of multiple measures that may factor into an educator's overall effectiveness rating. However, there are different ways that MGPs can be used for this purpose, and district decision makers will want to make informed choices to this end. The purpose of this brief is to provide school districts in Colorado with insights and key considerations in regard to the use of SGPs aggregated to the teacher-level in educator evaluations. Two broader topics addressed in this brief are: 1) the categorization of MGPs to distinguish teachers along the teacher MGP distribution; and 2) technical and policy design considerations for using and reporting MGPs for educator evaluations.

It is common for "cut-points" to be chosen to establish discrete categories within any distribution of teacher-level MGPs. The purpose of these categories is to draw distinctions among teachers whose students have demonstrated qualitatively distinct amounts of growth in achievement, on average. MGPs are typically categorized by states and districts into a minimum of three and a maximum of five discrete categories representing different levels of growth (e.g., low, typical and high growth). Two common approaches used to categorize MGPs in different states and districts are:

- A fixed-cuts approach.
- A fixed-cuts approach with confidence intervals (teacher-specific or a constant margin of error).

Each of these approaches is discussed in this brief. The second broader topic addressed in this brief relates to additional technical and policy design considerations for using and reporting teacher MGPs in educator evaluations. More specifically, the issues addressed under this topic are:

- Developing inclusion or eligibility rules.
- Using either the mean or median for teacher-level MGPs.
- Pooling SGPs across years.
- Combining the MGPs by subject to report an overall growth score on state assessments for a given teacher.
- Setting a minimum number of SGPs.
- Adjusting for possible sources of bias by factoring in teacher and student characteristics.
- Weighting the teacher-level MGPs.

The mean or median of Student Growth Percentiles, an MGP, is just one of multiple measures that may factor into an educator's overall effectiveness rating. However, there are different ways that MGPs can be used for this purpose, and district decision makers will want to make informed choices to this end.

We begin this brief with a general background on the CGM, and common approaches used to analyze the growth percentiles, before moving into the first topic of categorizing the MGPs. We then address each of the issues for the second topic in the same order as the points noted above, and conclude by emphasizing the importance of considering the role of this one measure as part of a larger system of measures used to inform an educator's evaluation.

Background on the Colorado Growth Model and Student Growth Percentiles in Educator Evaluations

The Colorado Growth Model (Betebenner, 2009) and its accompanying graphical interface (www.Schoolview.org) serve as the backbone of Colorado's approach to educational accountability. The Colorado Growth Model is used to compute a conditional test score percentile, known as a student growth percentile, for each student. This student growth percentile is found, in essence, by comparing the test score performance of a student in a given grade to all students in the state with a similar test score history in prior grades. A student that scores at the 50th percentile of this conditional distribution is one that is inferred to have shown "growth" that represents "one year of learning." Note that in this approach, the concept of growth is an inference—we infer that if a student has performance that is higher than expected relative to similar students, the reason for this is that they have learned more (i.e., shown more growth) than these similar students. An estimate of classroom or school-level growth can then be computed by taking the median or mean over all student growth percentiles for students with test scores in at least two adjacent grades. The result is what is known as an "MGP" (i.e., a median growth percentile or a mean growth percentile), which is the focus of the sections in this technical brief.

A recently released paper identified three common approaches used by states to aid in the interpretation of student growth percentiles for accountability and other purposes: norm-referenced growth, criterion-referenced growth, and baseline-referenced growth (Betebenner, Diaz-Bilello, Marion & Domaleski, 2014). Two of these approaches, the norm-referenced and baseline-referenced growth approaches, are commonly used in educator evaluation systems established across many states and districts. Please see FAQ 1 for a discussion of criterion-referenced growth.

FAQ 1 – Why shouldn't our district use criterion-referenced growth data (such as the catch-up and keep-up data provided by the state) for educator evaluations?

Currently, the criterion-referenced data (i.e., growth to standard data) for individual students are provided to all Colorado school districts in two ways: 1) identifying "unsatisfactory" and "partially proficient" students on the TCAP who are targeted to be on track to reaching proficient in 3 years, or by the time they reach grade 10; and 2) identifying "proficient" students on the TCAP who are targeted to maintain proficiency in 3 years, or by the time they reach grade 10. The first group of students is referred to in the state as "catch-up" students and the second group of students is referred to as "keep-up" students. Many districts in Colorado compute the percentage of catch-up and keep-up students each year to evaluate whether adequate growth is being made with students who are below proficient to reach proficiency, and with those students who have reached proficiency to maintain proficiency.

We recommend against the use of catch-up and keep-up data for educator evaluation purposes because it may create an uneven playing field. It is quite challenging for students scoring at the "unsatisfactory" level and the lower levels of the "partially proficient" score range to reach proficient within a three year time frame, therefore teachers instructing students at these proficiency levels would be disadvantaged by the catch-up metric. Conversely, because of the high likelihood of "proficient" students maintaining their proficiency level, teachers instructing students with high levels of proficiency will be advantaged by the keep-up metric. An additional implication of using these criterion-referenced growth data for an educator evaluation system is that although "unsatisfactory" or "partially proficient" students may have made considerable growth (i.e., higher than the 50th percentile) relative to their peer groups, their accomplishments would not be recognized by the growth to standard metric. For these reasons, we assert that incorporating criterion-referenced growth data would result in an uneven playing field where a teacher's score or rating on this state measure would be dictated largely by achievement status rather than by growth.

In both the norm-referenced and baseline-referenced approaches, SGPs are aggregated using either the median or the mean to report average growth achieved at the teacher level. A teacher's score for a given content area or for their overall growth rating on state assessments (i.e., for those teachers with only one content area attributed to them) is then based on the location of a teacher's MGP relative to the cut-points set in the MGP distribution. The key difference between the norm-referenced and baseline-referenced approaches is that under the latter approach, MGPs are interpreted relative to a baseline year of interest. In a purely norm-referenced approach, the SGPs computed each year for a state or school district are based on the most recent cohort of students taking state tests. For this reason, the median of SGPs for all students will always be 50—each year half of all students will have SGPs below average, and half will have SGPs above average. In contrast, a baseline-referenced approach compares the growth results from each year relative to a baseline year (or baseline years) of interest. Because of this, if more recent cohorts of students are showing more growth than previous cohorts, it would be possible to detect this over time. Under a baseline-referenced growth approach, a state's or school district's performance can either shift below or above 50 relative to the baseline year of interest. Although the baseline-referenced growth concept is appealing, it can be very challenging to implement and maintain because it depends upon maintaining the comparability of test scores from year to year (i.e., the psychometric process of "equating" test scores to adjust for random differences in difficulty).

After aggregating the SGPs to the teacher level using either the mean or the median, there are three common approaches taken by states and districts to make distinctions among teachers within any MGP distribution: scoring teacher MGPs using fixed-cuts, scoring teacher MGPs using fixed-cuts with teacher-specific confidence intervals, or scoring teachers using fixed-cuts and a margin of error. In the next section, we discuss each of these approaches used to categorize or make distinctions among teachers in the system based on the distribution of teacher-level MGPs.

Approaches for Categorizing MGPs

Fixed-Cuts Approach

The simplest approach that could be taken to categorize teachers according to their MGPs would be to score a teacher based on whether his or her MGP fell within established cut-points (please see FAQ 2 for information on how many cut-points or scoring groups should be set in the MGP distribution). For example, in Massachusetts, only three discrete score categories are used to distinguish among teacher MGPs: low, moderate and high growth. A rating of "low" is assigned to teachers with MGPs of 35 or lower. Teachers with MGPs greater than 35 but less than 65 earn a "moderate" rating and those teachers with MGPs of 65 and higher earn a "high" rating.

FAQ 2: How many scoring groups should be set for growth?

The number of scoring categories set in the MGP distribution depends on the desired number of scoring groups that a district would like to implement for growth, but with the understanding that enough distance between cut-points or a range of MGPs should be considered to effectively distinguish performance between teachers. In some places, such as Hawaii, the number of scoring groups for growth is based on the desire to maintain a parallel scoring structure across the system. That is, since there are four possible overall rating categories for the overall system, four scoring categories were set for growth and all other components in the system. For MGPs, we advise a minimum of 3 categories and a maximum of 5 categories since the range of MGPs falling under each scoring group becomes narrower with the addition of each category and more difficult to differentiate. To define the cut-points for varying levels of growth, criterion-referenced growth data can be reviewed to identify the level of growth required on average for below proficient students to move out of the unsatisfactory to the partially proficient category or for proficient students to move into the advanced categories. This information can then be used to help justify the placement of cuts in the MGP distribution (note that this approach does not have the same teacher-level implications as using criterion-referenced growth data as actual measures of teacher effectiveness). Other possible approaches to set cut-scores in the MGP distribution involve using quartiles if 4 scoring categories are desired or quintiles if 5 scoring categories are desired.

The simplicity of communicating results when MGPs are categorized using a fixed approach makes this method appealing in that it is straightforward to determine what score or rating any teacher will earn given an MGP value associated with student test performance in math, reading or writing. However, this approach ignores the influence of measurement error on each teacher's MGP. In other words, it does not take into account the fact that teachers with smaller numbers of students are more likely to be falsely placed into a lower scoring category when they in fact belong to a higher scoring category, or vice-versa.

Fixed-Cuts with Confidence Intervals

The key motivation for the use of confidence intervals in conjunction with the fixed-cuts to create score categories is that, in any given year, a teacher's observed MGP is a "noisy" estimate of his or her true MGP. To capture this idea, a confidence interval is computed and placed around each teacher's observed MGP or alternatively, the margin of error can be computed and placed around the fixed-cuts (please see FAQ 3) for more information about confidence intervals). For each teacher, the lower and upper bounds on plausible MGPs indicated by a confidence interval is reviewed relative to the cut-points set in the system to classify different levels of growth in the MGP distribution. Depending on the scoring rules used, either the upper and lower bound of the interval may be assessed relative to the cut-points set in the MGP distribution, or just one end of the tail may be assessed. We address both approaches in this section.

FAQ 3: What are confidence intervals?

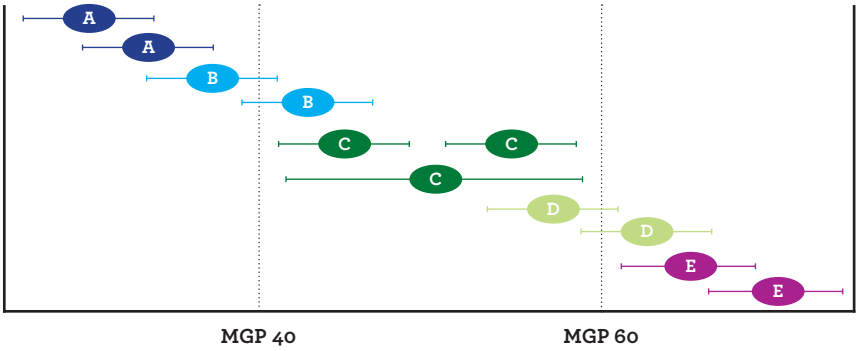
All teachers know that there is a component of chance that factors into their average student growth percentile each year. Students can be more challenging or less challenging from one year to next, which can influence a teacher's MGP. This means that in any single year, it is important to capture the uncertainty in a teacher's MGP that can be attributed to chance differences in students taught from year to year. This uncertainty is captured by forming a confidence interval around the MGP that is computed for each teacher in the current year. The confidence interval is used to quantify the lower and upper MGP values a teacher could have plausibly received if she had a higher or lower performing group of students. The width of a confidence interval depends on two things: the variability in student test performance and the number of students that were used to compute the teacher's MGP. A teacher with many students who don't vary much in their test performance will have a narrow confidence interval since the MGP can be more precisely estimated with more data points (i.e., having more students) and with similar performing students (i.e., less noise in the data). A teacher with few students who vary a lot in their test performance will have a wide confidence interval since there are fewer data points available to compute the teacher's MGP and the wider range of performance has been observed. The use of confidence intervals ensures that no teacher will be disadvantaged due to the year to year differences found in the type of students being instructed or because they worked with a particularly small group of students.

- **Fixed-Cuts with teacher-specific confidence intervals**

Figure 1 presents an example of how teacher-specific confidence intervals are applied in the educator evaluation system that has been developed at Denver Public Schools (DPS). In Figure 1, each of the dots represents a hypothetical teacher, and each dot is bounded by a confidence interval. The lower bound of the interval is represented by the line connected to the left of each dot and the upper bound of the confidence interval is represented by the line connected to the right of each dot. In DPS, a teacher can earn 5 possible scores (A-E, A indicating a low score and E indicating a high score) for growth on the state assessments depending on two factors: the MGP achieved by students, and the location of the upper and lower bound of the confidence interval relative to the low and high growth cut-points set in the MGP distribution. A score of A, C, or E is received if all three data points considered for a teacher – the MGP, the lower limit of the confidence interval, and the upper limit of the confidence interval – fall between the cut points. If one of the confidence intervals crosses a cut point, a teacher will fall into either category B or D.

Figure 1. MGP classifications and confidence intervals

(Source: Denver Public Schools)



Districts interested in this approach may also consider assigning scores based only on the upper bound of the interval. This alternative scoring rule would give teachers the "benefit of the doubt" and have the effect of moving more teachers out of the lower performing regions. This alternative scoring approach is used at the school level in places such as Colorado, Texas and Connecticut for No Child Left Behind, where the upper bound of the interval is only considered to give schools the benefit of the doubt for meeting proficiency targets.

From a technical standpoint, the use of teacher-specific confidence intervals makes it easier to assess the precision of each teacher's estimated MGP. The narrower the confidence interval, the more precise the estimate. However, one important communications challenge to consider with the teacher-specific confidence intervals is that two teachers with the same MGP value could receive different ratings depending upon the length of the confidence interval for each teacher. For example, if applying the decision rules used in DPS, two teachers with an MGP of 45 could receive a rating of a "B" or a "C" depending upon the length of the confidence intervals. If one teacher had a large group of students, a narrow interval, and the lower bound of the interval stays above an MGP of 40, this teacher would receive a classification of a "C". If the other teacher had a very small group of students, a wide interval, and the lower bound of the interval is lower than 40, then this teacher would receive a lower rating of a "B". However, in the case of the teacher with the smaller group of students, the wider interval should be expected since there is less information available to determine the true location of the MGP.

- **Fixed-cuts with a margin of error approach**

Another option that may be considered by a district is to use a constant margin of error with the fixed-cuts. The margin of error would apply to all teachers regardless of differences found in classroom composition among teachers (i.e., number of students associated with a teacher, and the variability of student performance). Rhode Island is planning to use this approach in their educator evaluation system. Under this approach, a margin of error is computed and established around each cut point in the MGP distribution (e.g., plus or minus 5 around each cut point to represent 1 standard error). All teachers with observed MGPs falling below the lower bound of the margin of error computed would fall into the lower rating category while teachers with observed MGPs above the higher bound of the margin of error would fall into the higher rating category.

Under this approach, two teachers with the same MGP value will always be evaluated in the same way relative to the margin of error set for each cut-point. Although the simplicity of communicating the results of this approach to teachers is appealing, this approach ignores the fact that classrooms often differ in composition, which can impact teachers' MGPs.

Having presented different approaches for consideration to categorize MGPs, we now turn to additional technical and policy areas related to the use and reporting of MGPs to make inferences about teacher effectiveness. These areas are:

- Developing inclusion or eligibility rules.
- Using either the mean or median for teacher-level MGPs.
- Pooling the SGPs data across years.
- Combining the MGPs by subject to report an overall growth score on state assessments for a given teacher.
- Setting a minimum number of SGPs.
- Adjusting for possible sources of bias in the growth data by factoring in teacher and student characteristics.
- Weighting the teacher-level MGPs.

Each of the areas noted above should be reviewed and considered by school districts using and reporting teacher-level MGPs in their educator evaluation system and will require that dedicated staff members are available to conduct analyses to help drive design choices in using the SGPs for accountability purposes. Although there is no specific order for districts to tackle each area addressed in the next section, we recommend that districts should consider all of these areas in order to demonstrate to their stakeholders that a deliberative and informed process has been followed to design the system.

Although there is no specific order for districts to tackle each area addressed in the next section, we recommend that districts should consider all of these areas in order to demonstrate to their stakeholders that a deliberative and informed process has been followed to design the system.

Additional Design Considerations for Computing and Interpreting MGPs

Inclusion/Eligibility Rules

Although inclusion rules are not commonly discussed in most reports published about using growth data to support educator evaluations, districts need to establish clear inclusion rules and also monitor the quality of student-teacher links since both are vital to providing quality and fair data for educator evaluation purposes. Roster verification processes deployed in many states and districts serve as a critical first step for determining which students belong to which teachers.

After verifying which students belong to which teachers, inclusion and eligibility rules are developed and applied with stakeholder input to determine which scores should count or be excluded from the evaluation of a teacher. Inclusion rules reflect local policy decisions that are made to help achieve fairness in an evaluation system. The standards set for inclusion rules will likely vary across districts based on stakeholder input received in each setting. We highlight inclusion rules from two districts in Colorado (Denver Public Schools and Harrison School District Two) as examples of the types of rules that have been established to contribute toward achieving fairness in their systems.

Inclusion rules reflect local policy decisions that are made to help achieve fairness in an evaluation system. The standards set for inclusion rules will likely vary across districts based on stakeholder input received in each setting.

In Denver Public Schools (DPS), the district requires a minimum of at least 80 percent contact time to be achieved by *both* teacher and student in order for a student's SGP to be counted in a teacher's evaluation. That is, a teacher must have instructed the student for 80 percent of the time associated with a course and the student must have enrolled in the classroom for 80 percent of the time associated with a course. If these two criteria are not met, the SGP for a student is not the linked to the teacher. Additional information about inclusion rules used in DPS is located in Attachment A.

Similar to DPS, Harrison also developed inclusion rules for both teachers and students to ensure that only test scores that meet different criteria can be linked to an educator. Whereas DPS uses the instructional period to determine which scores count in a teacher's evaluation, Harrison uses the assessment period to determine which scores count. In addition, Harrison also noted that scores from students who are "habitually absent" or who miss 25 percent or more of the assessment window are automatically excluded from an educator's evaluation. Additional information and details of eligibility rules used in Harrison can be found in Attachment B.

Mean vs. Median

There are two commonly invoked alternatives for summarizing the "average" of the SGPs associated with any given teacher. The first is to sum up all SGPs and then divide by the total number of students. This would result in a *mean* student growth percentile. The second is to order the SGPs from lowest to highest and then pick the value that splits the list in half (i.e., the 50th percentile). This would result in a *median* student growth percentile.

The median is typically used with percentiles because percentiles are not traditionally deemed to be on an interval scale. That is, the difference between students located at the 4th and the 5th percentile is not considered to be the same as the difference between students located at the 53rd and 54th percentile since percentile differences can represent different intervals of actual points achieved by students at those locations despite being separated by the same gap. However, if the list of SGPs is normally distributed (i.e., bell-shaped) then the mean and median will be the same. If the SGPs are not normally distributed, the mean and median will usually differ. Because the mean is sensitive to outliers and the median is not, the difference between a mean and median SGP will tend to be most pronounced for teachers with fewer students. In general, the choice of mean vs. median will interact with the approach taken to place teachers into score categories (see previous section). If an approach that only uses cut-points without confidence intervals is being taken, the median will generally be the most defensible option. If a cut-point with confidence interval approach is being used, the mean is the preferred option (Castellano & Ho, 2013).

Pooling SGPs across Years

As discussed earlier in reference to the use of confidence intervals, MGPs are sensitive to sample size (see McCaffrey et al., 2009). One way to help mitigate the role of chance error (see FAQ 3) is to pool SGPs across multiple cohorts (i.e., years) of students. That is, by increasing the number of students included in a teacher's evaluation, there is more information available to compute a teacher's MGP. For example, in the DPS educator evaluation system, a teacher's MGP is computed using up to three years of data, if available. When this is done in conjunction with the computation of confidence intervals, the confidence intervals using three years of data were always found to be narrower than the width of the intervals found for the MGPs when using just one year of data. The confidence interval for an MGP pooled across multiple cohorts or years will always be narrower than the confidence interval for an MGP based on a single cohort of students, simply because the number of scores available to compute the MGP will be larger. One area to be considered is that the decision to pool data means that specific year-to-year MGP changes in performance will be masked through this approach since the average performance achieved across years is being reported.

The confidence interval for an MGP pooled across multiple cohorts or years will always be narrower than the confidence interval for an MGP based on a single cohort of students, simply because the number of scores available to compute the MGP will be larger.

Combining MGP Scores for Teachers with Multiple Student Outcomes

For teachers with students taking multiple state assessments, one can imagine a possible maximum of 3 MGP score categorizations, representing scores for reading, writing, and math. In some states (e.g., New Jersey, New York, Hawaii and Georgia) the SGPs across content areas are combined and only one MGP is reported for each teacher. However, Colorado's Technical Advisory Panel for Growth has recommended that the SGPs for each content area should be summarized as separate MGPs in order to preserve the information obtained and reported for each content area to teachers. This recommendation is consistent with how the MGPs are reported and scored separately by content area for school and district accountability in the state prior to aggregating the score to report a final growth rating. For some teachers, the reporting of separate MGPs would entail having to combine the results from multiple MGPs to arrive at a final categorization.

Two possible approaches for combining scores across content areas are to use a decision matrix or to establish a composite index score. Under the decision matrix approach, a 2 by 2 matrix with scores from two content areas may be considered and the cells of the matrix are populated using values agreed upon by stakeholders. Figure 2 illustrates an example of a decision matrix in the case of a teacher with only two content areas being assessed. A teacher's final score on state assessments is based on the rules defined by the matrix in the shaded cells. In this example, if a teacher achieves a 4 on reading and achieves a 2 on writing, the overall score for this teacher is a 3. However, if a teacher has 3 content areas, and the goal is to equally weight the information from all three subjects, then this approach would not be recommended. But if a district wants to consider the scores from reading and writing as a single "English language arts" score, then the scores from combining reading and writing may be established first, and then compared against the score from mathematics to determine the final overall score for growth. When the state transitions into the new assessments in 2014-2015, the issue of combining three content areas using a decision matrix will no longer be a consideration since there will only be scores available from two content areas.

Figure 2. Decision Matrix Example

Score from TCAP Writing	Score from TCAP Reading			
	1	2	3	4
1	1	1	2	2
2	1	2	2	3
3	2	2	3	3
4	2	3	3	4

When using a composite index approach, different weighting considerations (i.e., equal across all content areas or weighting scores more heavily for one or two areas) may be considered prior to aggregating the scores earned across content areas using the mean or the sum of scores. For example, if the mean is used to aggregate the score, the composite index score for a teacher who earned a 2 on writing, 1 on math, and 3 on reading is 2. If the sum of scores is used, cut-points are then set within the total possible points earned to establish the final overall score earned by teachers with 2 or 3 scores available. Using the same results from the earlier example, the sum of scores earned by this same teacher is 6. A district may set the cut-points for teachers with 3 content area scores as follows: 1-3 points = 1; 4-5 points = 2; 6-7 points = 3; and 8-9 points = 4. Based on these score ranges, the teacher's final score in this example is a 3. Additional discussion on the topic of combining scores can also be found in CDE's educator effectiveness website[1].

Setting a Minimum Number of SGPs

Most districts will establish business rules for the minimum number of students with valid SGPs necessary before it is reasonable to compute and report an MGP for a teacher. Typical values for this minimum number found across states range from 10 to 20. The choice of minimum number of SGPs can interact with the decision to pool across multiple cohorts. For example, a district might choose to compute an MGP for a teacher so long as the cumulative sample size reaches some minimum threshold. In the DPS educator evaluation system, an MGP is computed for a teacher so long as that teacher has at least 15 students with SGPs over three years, and at least 5 students in the most recent year. The choice of minimum number of SGPs also interacts with the decision to classify teachers using a confidence interval. If no confidence interval is being used, it will generally be sensible to specify a more conservative minimum number. However, even if a confidence interval is being used, the computation of an MGP on the basis of fewer than 10 students is unadvisable.

Adjusting for Possible Sources of Bias

In the Colorado Growth Model, an SGP is computed by comparing each student's achievement to that of peers with similar test score histories in prior grades. In this sense an MGP can level the playing field for teachers who have predominantly low-achieving students relative to teachers who have predominantly high-achieving students. However, the Colorado Growth Model does not adjust statistically for other variables outside a teacher's control that might contribute to a student performing better or worse than expected on a standardized test. For example, teachers with higher proportions of English Language Learners or students with individualized education plans may face unique challenges that are not captured by the Colorado Growth Model. A key consideration for many school districts and states is the correlation between teacher MGPs and contextual variables such as the percent of students who are English Language Learners, percent in special education, percent eligible for free and reduced price lunches, etc. The correlations provide a measure of how strongly associated the level of growth achieved by a teacher's students relative to each contextual variable of interest. For example, if the correlation between the MGPs and English Language Learners are negatively strong (e.g., -.8 and above), this finding would suggest that teachers with a high percentage of English Language Learners are likely to have low MGPs and thereby would also likely score low on this one measure of their overall evaluation.

Although research has found that the growth results reported in the state's performance frameworks are weakly correlated with student demographics[2], the strength of the correlations will vary by district. If the correlation between growth and contextual variables of interest is found to be moderate to strong in a district, it can indicate that the most disadvantaged students are receiving lower quality instruction (unfair to students scenario), or that the MGP is biased by these factors (unfair to teachers scenario). If the latter is of greatest concern it would be possible to further adjust MGPs by comparing subsets of teachers with the same classroom contexts. For example, teachers in a school district could be stratified by the percentage of English Language Learners with subsequent MGP comparisons taking place within a stratum. A more complex version of this approach would compute adjusted MGPs by regressing the observed MGPs of teachers onto classroom context variables. For each teacher, an adjusted MGP would be the observed MGP minus the predicted MGP, with adjusted MGPs having a mean of 0.

1. See Determining a Final Educator Effectiveness Rating: http://www.cde.state.co.us/sites/default/files/Determining%20Final%20Rating%20TEACHER_Revised_2014_01_14.pdf

2. See A Demographic Review of School and District Performance Framework Outcomes: http://www.cde.state.co.us/sites/default/files/Demographic%20Review%20Paper.pdf

Although there is considerable appeal in adjusting MGPs for additional contextual factors, there are at least two important tradeoffs to navigate. First, making secondary adjustments to MGPs can impose serious data management and analysis burdens on a school district. An MGP is a simple statistic to compute from the SGPs provided directly by the state; an adjusted MGP must be computed by a capable secondary analyst. The more contextual factors for which an adjustment is desired, the greater the chance for a mistake to be made in the attribution of variables to students and their teachers of record (e.g., multiple students labeled as English Language Learner after they have been reclassified with full English proficiency). Second, by including additional factors in a secondary adjustment, it becomes possible to ensure that teacher MGPs and contextual variables are uncorrelated by definition. Yet if, in fact, some teachers are more effective with certain kinds of classrooms, then such an approach will tend to *overadjust* MGPs. It follows that the decision to incorporate a secondary adjustment on MGPs should not be taken lightly.

By including additional factors in a secondary adjustment, it becomes possible to ensure that teacher MGPs and contextual variables are uncorrelated by definition. Yet if, in fact, some teachers are more effective with certain kinds of classrooms, then such an approach will tend to overadjust MGPs. It follows that the decision to incorporate a secondary adjustment on MGPs should not be taken lightly.

Although the decision to include or omit additional contextual variables when comparing teacher MGPs is well-understood as an important issue that can serve to reduce or increase bias, the impact of measurement error on MGP bias has only recently been given the same degree of attention. In short, the Colorado Growth Model is at heart a sophisticated version of a classic regression model. Such models typically assume that an outcome variable is being predicted by independent variables that have no measurement error. This assumption is obviously implausible when a student's prior year test score is the independent variable in question. If students are assigned to teachers within schools throughout a district or state as if at random, and if they vary considerably in prior achievement by classroom, then the impact of measurement error on MGPs will be negligible. If students are sorted into certain kinds of classrooms as a function of their prior achievement, the impact of measurement error on the MGPs of teachers in the lowest and highest achieving classrooms can be substantial (i.e., teachers in low achieving classrooms will have MGPs that are too low; teachers in high achieving classrooms will have MGPs that are too high). In Georgia's educator effectiveness system, an approach that relies on what is known as a SIMEX correction (see Shang, 2012) has been developed to adjust for measurement error in teacher and school-level MGPs. However, the approach is computational intensive and may be outside the scope of resources of most school districts to implement.

Weighted MGPs

In some states (e.g., New York and Hawaii), the MGPs reported for each teacher reflect a weighted mean or median. That is, the MGP computed for each teacher is adjusted based on a combination of a student's enrollment and attendance rate in the class. This approach is used in some educator evaluation systems to ensure that students with lower attendance and enrollment rates contribute less toward the computation of a teacher-level MGP relative to students who have higher attendance and enrollment rates. If all students associated with a given teacher attended for 100 percent of the enrollment period and also attended for 100 percent of the attendance period, then the MGP would be the same as the unweighted MGP since no adjustments would need to be made to the individual SGPs. Adjustments are only made to the individual SGPs when enrollment and attendance rates vary across students. Table 1 presents an example of how a weighted MGP is calculated for a hypothetical classroom of three students with varying enrollment and attendance percentages.

Table 1. Hypothetical data used to develop weighted MGPs

	SGP	Percentage of time enrolled in class	Percentage of time attended class	Weight (enrollment * attendance
Student A	56	80%	90%	.72
Student B	73	100%	75%	.75
Student C	35	100%	95%	.95

Using the method applied in New York, computing the weighted average for the results in Table 1 would entail undergoing the following steps:

1. Take the sum of the weighted scores across all students: (56 *.72) + (73 *.75) + (35 *.95) = **128.32**
2. Take the sum of the weights: .72 + .75 +.95 = **2.42**
3. Divide the weighted score (from step 1) by the sum of the weights (from step 2) to calculate the weighted MGP: 128.32/2.42 = **53**

For this hypothetical classroom with 3 students, the weighted MGP would equal 53. The unweighted mean growth percentile for this classroom is 55 and the unweighted median growth percentile is 56. In this example, the weighted MGP reported is dampened by the student with the SGP of 35 who had a higher weight (equated to an attendance by enrollment rate) compared to his peers.

This decision to weight the MGPs should be made with stakeholder input since this reflects a policy decision. That is, this decision is based on a shared belief on the extent to which a teacher's MGPs should be weighted to reflect differences in each student's instructional interaction time with the teacher. One additional consideration for districts interested in using this approach is that a strong data quality assurance process must be established in order to gather accurate enrollment and attendance data on each student and to accurately compute the weighted MGPs for all teachers.

Conclusion

Since Colorado districts have considerable flexibility in deciding how to use the MGPs for educator evaluations, the primary objective of this brief is to help inform and highlight technical and policy areas that should be reviewed and considered by districts using teacher-level MGPs in their evaluation. We recommend that districts wishing to incorporate teacher-level MGPs in their educator evaluation system dedicate time and resources to conduct analyses in order to assess how results vary based on the different design choices made to both categorize and apply different approaches to using the MGPs to inform educator evaluations. For example, a district may find that after pooling data across different years, that pooling across more than three years of data does not add much to the stability or reliability of results found from year to year. This analysis may lead a district to restrict the total number of years used to pool data to three years.

One area for future consideration by all districts is the use of median growth percentiles during the assessment transition period, specifically for the 2014-2015 school year (or during the 2015-2016 school year for districts that are using prior year scores). Transitional median growth percentiles will still be calculated based on 2015 state assessment data, but they will not be included within the 2015 school and district performance frameworks, for reasons pertaining to the timing of the release of data and the need for comparability analysis.[3] Additionally, the transitional student growth percentiles may not be available until the winter of 2016. Districts wishing to incorporate the transitional MGPs from the new state assessments into the 2014-2015 evaluation cycle (or the 2015-2016 evaluation cycle if using prior year scores) should weigh their decision against these factors.

Districts will need to consider how best to balance the contribution of data from each measure to inform an overall effectiveness rating. MGPs are most likely to be useful as a basis for evaluating teachers when this information is properly balanced against other information about student performance (e.g., evidence from student learning objectives) and direct observations of teaching practice.

Finally, an important point to reiterate from the introduction of this brief is that the MGPs consist of only one of multiple measures being used to evaluate teachers. Districts will need to consider how best to balance the contribution of data from each measure to inform an overall effectiveness rating. MGPs are most likely to be useful as a basis for evaluating teachers when this information is properly balanced against other information about student performance (e.g., evidence from student learning objectives) and direct observations of teaching practice. Further when one measure does not impose an undue influence on the system, this allows for stakeholders to appreciate how all measures are used together to help inform an overall judgment about teachers. Under this scenario, an MGP, like any other measure in the system serves as one of many sources of information that are evaluated using human judgment about teaching quality in order to make informed personnel decisions. This perspective is also consistent with one of the recommendations made to the state by the State Council for Educator Effectiveness advising on the design and implementation of the state's educator evaluation model. As noted by the State Council (2011), "Data should inform decisions, but human judgment will always be an essential component of [educator] evaluations" (p. 6).

3. See Supporting the Accountability Transition (Draft Plan):
http://www.cde.state.co.us/sites/default/files/Accountability%20Transition%20Fact%20Sheet%201-29-2014%20FINAL.pdf

References

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

Betebenner, D.W., Diaz-Bilello, E.K., Marion, S.F., Domaleski, C. (2014). Using student growth percentiles during assessment transition: technical, practical and political implications. Retrieved from www.nciea.org.

Castellano, K. E., & Ho, A. D. (2013). Practical differences among aggregate-level conditional status metrics: From median Student Growth Percentiles to value-added models. Retrieved from: http://scholar.harvard.edu/files/andrewho/files/practical_differences_among_acsms_-_castellano_and_ho_2013.pdf

McCaffrey, D. F, Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4), 572–606.

Shang, Y. (2012). Measurement Error Adjustment Using the SIMEX Method: An Application to Student Growth Percentiles. Journal of Educational Measurement, 49: 446–465.

State Council for Educator Effectiveness (2011). State Council for Educator Effectiveness Report and Recommendations. Retrieved from www.cde.state.co.us.

Center for Assessment

National Center for the Improvement of Educational Assessment (NCIEA)
Dover, New Hampshire

CADRE

Center for Assessment Design Research and Evaluation (CADRE)
University of Colorado, Boulder

Attachment A

Student-Teacher Attribution in Denver Public Schools: Whose Scores Are Included in Educator Evaluations?

One of the most important considerations for using teacher level growth percentiles in teacher evaluation systems is determining how to identify the appropriate student growth percentiles to include in these calculations. As DPS has implemented teacher-level growth percentile scores in the Leading Effective Academic Practice (LEAP) system, we've received more questions about this from teachers and school leaders than about any other aspect of these scores.

At issue here is student-teacher contact time, or the amount of instructional time a student and teacher spend together. Students frequently transfer to different schools during a year and move into different course sections mid-course. Teachers also commonly move to different schools during a school year, or go on leave. This movement affects the time that a teacher and student spend together, which, in turn, affects the impact the teacher is able to have on the student's learning.

To account for this, DPS uses rules and processes for identifying the pools of student growth percentiles on which teachers' evaluation results are based. In developing these rules and processes for LEAP, we looked first to our experience with the Exceeds Expectations program in DPS's Professional Compensation (ProComp) system for teachers. Under Exceeds Expectations, teachers receive bonuses if at least 50% of the students assigned to them in any given school year have student growth percentiles in reading, writing, or math of 55 or higher. Exceeds Expectations specifies that students must be enrolled in a reading, writing, or math course for at least 85% of the course duration, and in attendance for at least 85% of the days they're enrolled.

To create student-teacher attribution rules for LEAP, we drew on the existing rules for Exceeds Expectations, but we were able to utilize our newly developed Teacher-Student Data Link (TSDL) system to create more precise ways to measure student-teacher contact time. TSDL allows us to identify the number of days (minutes, actually!) that a teacher and student are together in a course, based on the student's attendance and the teacher's assignment to the course. Using this information, we developed a simple rule that a student's growth percentile is only included in a teacher's pool if the teacher and student were together in a course for at least 80% of the course duration. While this this threshold may appear lower than the criteria for Exceeds Expectations, it is actually higher. A student growth percentile is included in Exceeds Expectations if a student is enrolled for 85% of a course and in attendance for 85% of the time enrolled. If a student meets both of these thresholds exactly, his or her growth percentile can be included if the student is in attendance for only 72.5% of the course. For the higher stakes context of LEAP, DPS wanted to establish more stringent criteria for student-teacher contact time. This rule applies to all courses linked to content areas of reading, writing, and math in the TSDL system. Using this rule and the TSDL system, we are able to filter out growth percentiles of students who spent insufficient time with teachers, giving us much greater confidence in the validity of our scores.

Lessons Learned

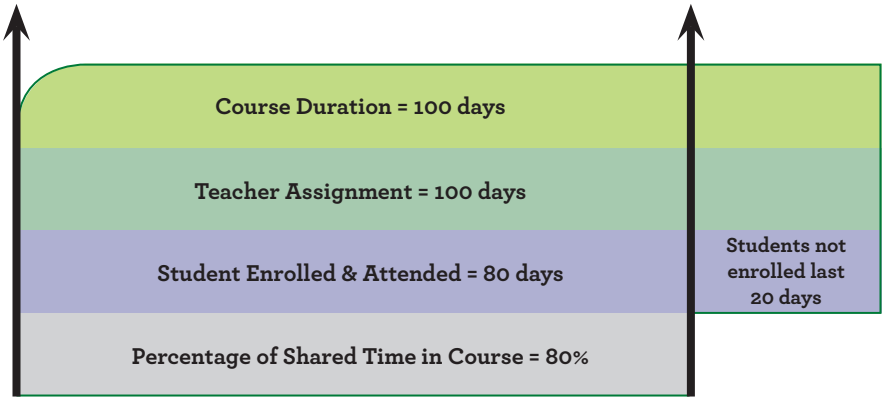
How can other districts go about establishing similar student-teacher attribution rules? A data system that links teachers, students, courses, and assessments greatly facilitates the process and enhances the accuracy of the rules. However, such systems can be complex and expensive to develop, and it's certainly possible to use attribution rules without complex data systems as long as information on student and teacher course assignments and student attendance by course are available. In DPS, we currently apply daily attendance, but we are examining the possibility of recording attendance by subject matter block to ensure greater accuracy. In elementary schools, attendance is typically taken only once per day, rather than for each "course" (e.g., reading, writing, math). Daily attendance can be applied to all subject areas, or attendance practices can be altered so that attendance is recorded for each subject matter block. Rules can be structured similarly to those for Exceeds Expectations, where enrollment and attendance criteria are examined separately, which simplifies analysis. Alternatively, rosters of eligible students for each teacher can be created by schools, using guidelines for student enrollment and attendance, eligible courses, and teacher assignments. A school-based reporting system can be much simpler to implement, as long as strong guidelines are in place and an auditing process is established to ensure the accuracy of the rosters.

More importantly, districts must establish criteria for student-teacher contact time that aligns with their expectations for teachers and students, and with their goals for their evaluation systems. In doing so, consider the following questions:

- What does instructional contact mean to us? Seat time in a course? Do we account for any non-instructional time that occurs during a course?
- How much instructional contact time is sufficient to reasonably expect a teacher to affect a student's academic growth?
- What are our expectations for student attendance and the time students should be enrolled in courses in order to see academic growth?
- What are our expectations for students' academic growth?

DPS had numerous conversations with teachers, school leaders, and central office school support staff to answer these questions. In doing so, we've come to agreements about expectations that carry over into other aspects of instructional improvement. At the same time, we consider this work to be a process of continuous improvement. As we implement LEAP and gather feedback from our stakeholders, we'll continue to examine our system and make adjustments to strengthen it so that our results are fair and accurate for our teachers, and effective in improving outcomes for our students.

Figure 1. Graphic of Attribution Rules



Attachment B

Eligibility Requirements for Staff and Accountability of Student Scores in Harrison School District Two's Effectiveness and Results Pay for Performance Plan

The following documents the eligibility criteria for staff and accountability of the student scores in Harrison School District Two's (HSD2) Effectiveness and Results (E&R) Pay for Performance Plan. The term "evaluation period" is used to refer to the year that the achievement data is being collected.

Staff Eligibility

In order to participate in the E&R Pay for Performance Plan, HSD2 teachers must abide by the following eligibility criteria for the results year(s):

1. The staff member was employed by HSD2.
2. The staff member must have physically worked 2/3 of their scheduled calendar.
3. At least 3/5ths of the staff member's courses taught fall under an existing Student Achievement Template. The 3/5ths rule applies to each semester for secondary staff. (i.e. a staff member must have 3/5ths of the courses taught under an existing Student Achievement Template for both the first semester and the second semester).
4. A course must have at least 12 students enrolled to qualify for the 3/5ths rule. The exceptions are SPED, CLDE, and AP courses where the enrollment may go below 12 students with the approval of the applicable department.
5. Staff members that have Student Teachers or Interns are accountable to the students instructed by their Student Teacher or Intern.
6. Reference *The Harrison Pay for Performance Plan* for details on staff on remediation/improvement plans or staff with unresolved legal issues.

For staff to be accountable for an assessment series (e.g. District Assessments, Specials Projects, etc.) they must not be on leave more than 50% or more of the student score eligibility period. (Reference Table 1)

Student Score Eligibility

Student score eligibility is based on the student's enrollment (specifically entry and withdrawal dates in the HSD2 Student Administration System) for the assessment period. It is the ethical responsibility of the school staff to ensure that all entry and withdrawal dates are accurate.

1. The student must be enrolled continuously in the accountable teacher's class(es) as defined by the HSD2 Official Instructional Calendar and school regulations for attendance.
2. At least one teacher must be accountable to every score if the student has been enrolled at the same school for the accountable assessment period.
3. Homebound students will not be accountable to any staff member.
4. CoAlt students will not be accountable to regular educational or moderate needs staff members.
5. Significant Support Needs (SSN) students, as identified by the SPED Department, will not be accountable to staff for their Art, Music, and PE assessment scores.

The above list represents a partial representation of all ineligibility reasons. The assessment eligibility period is defined by Table 1.

Table 1. Student Score Eligibility

Student Eligibility Criteria	
Assessment	**Continuously enrolled in the staff members class:**
Elementary Specials District Assessments	75% of the instructional block
Elementary Specials Performance Task	
District Assessments (including Secondary Electives)	75% of the instructional block
Secondary Electives Semester Projects	
End of Course Exams	
Elementary Specials Final Project/Performance	75% of the instructional block
3rd Gr Reading TCAP	Oct 1 – 1st day of assessment window
TCAP (MGP), ACCESS	Elementary: Oct 1 – 1st day of assessment window Secondary: Oct 1 – last day of Semester 1 OR 1st Monday of Semester 2 to assessment window
Accuplacer	
ACT	

Habitually Absent

The term habitually absent was defined by the E&R Focus Group. The E&R Focus Group is comprised of representatives from each school (teachers and administrators), School Supervision & Leadership, Research, Data & Accountability, Curriculum, Instruction & Assessment, SPED, and English Language Development (CLDE). The term habitually absent is deemed as:

> Student scores are not accountable to the assigned staff member if the student has been absent for 25% or more of the assessment eligibility window.

The 25% is applied at the class level, for each assessment. Elementary morning (am) and afternoon (pm) absences are counted as one day of absence each.

Absence Reasons

The codings counted as an absence in the Infinite Campus school administration system are:

Table 2. Absence Reasons

Reason	Description
E	Excused Absence
L	Parent Excused
PD	Prof Documentation
R	Runaway
S	Out of School Suspension
U	Unexcused Absence

Student Ineligibility Reasons

Students are not accountable to a staff member if they qualify for one of the following:

- Habitually Absent – student qualifies under guidelines above
- School Entry date – student score does not qualify as per Table 1
- Class Entry date – student score does not qualify as per Table 1
- Roster Removal – student was approved to be removed during the Roster Validation process
- CoAlt (aka CSAP-A) eligible (with the exception ofCoAlt students on the SPED_SSN template)

Snow Days

District approved snow days are removed from the calculation for habitually absent students. District approved days are not counted as days a student was present or absent from school.

Participation Rate

All students that are currently enrolled during the assessment window are required to take the accountable assessments as documented in the respective content Student Achievement Template. Students are required to take their course/grade appropriate assessments, even if they do not meet the eligibility requirements to be accountable to a staff member.

"After removing the scores of habitually absent students and the scores of those students who are not eligible to be counted on a given assessment, the threshold for participation will be at least 90%. In other words, the teacher must have recorded scores for at least 90% of the available data points once the scores of habitually absent and ineligible students are removed. If the teacher does not meet this threshold, he will be ineligible for promotion to the next level. He could, however, be rated at a lower level should the data support such an evaluation." (The Harrison Pay for Performance Plan)