CADRE

NOVEMBER 2020

Comparison of 2019 Cohort and Baseline Student Growth Percentiles

Benjamin R. Shear

A report prepared by the Center for Assessment, Design, Research and Evaluation (CADRE) at the CU Boulder School of Education.

School of Education
UNIVERSITY OF COLORADO BOULDER

Acknowledgements

The author would like to acknowledge helpful feedback on prior drafts of this report from Elena Diaz-Bilello, Marie Huchton, and Derek Briggs, as well as data analysis and research assistance from Sandy Student, Adam Van Iwaarden, and Damian Betebenner. The views expressed in the paper are solely those of the author and any errors are attributable to the author.

About CADRE

The Center for Assessment, Design, Research and Evaluation (CADRE) is housed in the School of Education at the University of Colorado Boulder. The mission of CADRE is to produce generalizable knowledge that improves the ability to assess student learning and to evaluate programs and methods that may have an effect on this learning. Projects undertaken by CADRE staff represent a collaboration with the ongoing activities in the School of Education, the University, and the broader national and international community of scholars and stakeholders involved in educational assessment and evaluation.

Suggested Citation

Shear, B.R. (2020). Comparison of 2019 Cohort and Baseline Student Growth Percentiles. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE). https://www.colorado.edu/cadre/node/373/attachment

Please direct any questions about this project to: benjamin.shear@colorado.edu

Table of Contents

Purpose

Each year, as required by state and federal statute, the Colorado Department of Education (CDE) produces accountability ratings for each school. At the Elementary and Middle school levels, these accountability ratings are based on two aspects of student performance on the state Colorado Measures of Academic Success (CMAS) tests - status and growth. Status is measured by the average scale scores students receive. Growth is measured by the quantile-regression based student growth percentile (SGP) model (Betebenner, 2009). Both metrics are also disaggregated by student subgroups. Although state statute requires measuring growth using the SGP model, there are multiple ways the model can be estimated. This report focuses on one particular technical choice CDE must make when deciding which version of the SGP model to use in measuring student growth, namely whether these measures should be based on so-called cohort or baseline-referenced SGPs. The purpose of this report is to:

1. Compare and contrast the interpretation of baseline-referenced versus cohort-referenced SGPs.

2. Evaluate empirically how much inferences about student or school-level growth results might differ if baseline-referenced SGPs were used in 2019 instead of cohort-referenced SGPs.

3. Discuss issues to consider when comparing baseline and cohort-referenced SGPs.

The analyses in this report were planned and carried out using data collected prior to the disruptions caused by COVID-19. This report is intended to provide background context and results relevant to the use of SGP data under standard test administration and reporting conditions and does not address concerns specific to the educational disruptions and challenges that have occurred during the pandemic.

Background

Although nearly half of all states currently use SGPs in their accountability systems (Data Quality Campaign, 2019), there are a number of statistical and technical choices that can be made when estimating SGPs that have received relatively little attention in the research literature. As a result, many practitioners may not be aware of the different technical choices and the practical consequences that such choices can make for the use and interpretation of student growth data reported in the form of SGPs. For states using quantile-regression-based SGPs to measure student growth, there are a number of practical choices that need to be considered including whether to summarize aggregate SGPs using the arithmetic mean or median, how many prior year scores to include, whether to adjust for test score measurement error, and whether to norm SGPs relative to a stable baseline cohort or an annually updated cohort. This report takes up the question of whether to norm SGPs relative to a stable baseline cohort or to an annually updated cohort; an accompanying report examines the use of corrections for test score measurement error.

Use of SGPs in Colorado School Accountability Ratings

The median SGP (MGP) of students enrolled at each school are a key component used to calculate school accountability ratings in Colorado's School Performance Framework (SPF).[1] A full description of the SPF system is beyond the scope of this document, but we note a few relevant aspects of the system here. For Elementary and Middle schools enrolling students in 3rd-5th grade or 6th-8th grade, respectively, 60% of the SPF rating is based on MGPs while 40% is based on students' average test scores. Schools earn a variable number of points depending upon whether the MGP for students overall or within certain demographic subgroups is above a fixed set of thresholds; the same is true for average test scores. These points are then combined in a formula that produces the final SPF rating. The logic for including MGP results alongside average test scores is that while average test scores provide information about one valued outcome (the level of achievement students have reached at the end of each grade), MGPs provide information about the amount of progress students made during the most recent year while enrolled in the school. MGPs are intended to provide an indicator of learning that is more sensitive to school instructional practices and policies. In addition, because students with high or low prior test scores can achieve high SGPs, MGPs are viewed as a fairer way to evaluate schools that may enroll students who initially enter that school with different levels of prior achievement.

Definition and Interpretation of Baseline-Referenced and Cohort-Referenced SGPs

SGPs are norm-referenced statistics, describing each student's current year scale score relative to other students with similar prior scale scores. Because they are norm-referenced, each student's SGP could differ depending upon the norming sample to which they are compared. The CDE website describes SGPs as telling us, "how a student's current test score compares with that of other similar students (students across the state whose previous test scores are similar). This process can be understood as a comparison to members of a student's academic peer group."[2] As a result, a student's SGP can vary depending upon how this "academic peer group" is defined.

There are two common approaches to defining the academic peer group, referred to as "cohort-referenced" or "baseline-referenced" SGPs (Betebenner et al., 2014). In the cohort-referenced SGP model, each student is compared to students in their same grade and year with similar prior scores; this comparison group changes each year. In the baseline-referenced model, each student is compared to students from a stable baseline cohort in the same grade with similar prior scores; this comparison group remains the same across years. In the cohort-referenced model, if two students have identical 3rd and 4th grade math test scores, but one student took the tests in 2017/2018 and the other took the tests in 2018/2019, these two students could have different SGPs because they would be compared to different norm groups. In the baseline-referenced model, however, both students would receive identical SGPs because they are compared to the same norm group. Put differently, cohort-referenced SGPs answer the question, "how does each student's current achievement compare to academic peers with similar prior scores in the student's own cohort?" while the baseline-referenced SGPs answer the question, "how does each student's current achievement compare to academic peers with similar prior scores in the baseline cohort?"

[1]See additional information here: https://www.cde.state.co.us/accountability/performanceframeworks.
[2]See: http://www.cde.state.co.us/accountability/growthmodelfaqs-general.

Why might baseline SGPs be preferred to cohort SGPs?

The primary reason to use baseline SGPs rather than cohort-referenced SGPs is to track changes in aggregate student growth over time. By construction, the statewide mean and median cohort-referenced SGP will be exactly 50 each year, even if student growth at the statewide level is changing over time. As a hypothetical example, imagine that all districts in the state decide to use a new 4th-grade curriculum, and that this curriculum more effectively helps students learn 4th-grade content, so that all 4th grade students make more progress, on average, than did 4th grade students in the previous year. Under the cohort-referenced model, the statewide average or median SGP statewide would remain at 50 and appear to indicate no difference in student growth. The average baseline SGP, on the other hand, could be higher than 50, indicating that by the end of 4th grade, at the state level, students this year attained higher scale scores than did students with similar 3rd grade scores in prior years. Baseline SGPs thus allow the state to track trends in student growth across years in a way that is not possible with cohort SGPs.

Although the distinction between baseline and cohort MGPs at the school level is not as straightforward, a similar rationale applies. Unlike the statewide average cohort-referenced SGP, which will be exactly 50 each year, the average or median cohort-referenced SGPs for individual schools or districts can be higher or lower than 50 in any given year, so that changes in growth can be detected regardless of which type of SGP is used. Thus, schools could observe increasing or decreasing trends in MGPs across years whether using cohort or baseline MGPs. However, if schools are attempting to understand the cause of increasing or decreasing MGPs across years, the use of baseline SGPs could simplify this task. Suppose students enrolled in a particular school this year have similar prior year scores to students enrolled the previous year, but the overall cohort MGP of students in the current year is higher than in the prior year. Aside from statistical uncertainty, the higher MGP could be due to either of two factors (or a combination of both): 1) students this year had higher growth or, 2) the statewide cohort of students used to compute the cohort SGPs differed. On the other hand, the use of baseline SGPs rules out the latter possibility, thus slightly simplifying the task of trying to interpret and understand changes in MGPs across years. In both cases, a complete understanding of the school's MGP results requires considering the growth metrics in context and alongside other data about school and community factors as well as average levels of current and prior achievement.

From an accountability standpoint, there is an additional rationale for choosing between baseline and cohort SGPs. Some stakeholders have expressed concern that cohort-referenced SGPs result in a "zero sum game" in which some schools and students will always have "low" SGPs or MGPs each year (Betebenner et al., 2014). In contrast, under the baseline SGP model, there is no limitation on the number of schools or students with SGPs or MGPs above or below 50 in a given year. Despite this intuition, it is not obvious how the use of baseline SGPs in place of cohort SGPs would actually affect school accountability ratings. Because both baseline and cohort SGP calculations are conditional on prior achievement, school-level MGPs using both methods are likely to be very highly correlated. That is, schools where students had the highest achievement relative to "academic peers" in the current cohort are likely to be the same schools with the highest achievement relative to "academic peers" in the baseline cohort. As a result,

using cohort or baseline SGPs would likely identify the same schools as having the highest or lowest growth scores. On the other hand, because points in the SPF system are awarded based upon whether a school's MGP reaches certain absolute thresholds, it is possible that while the rank ordering of schools is similar, more (or fewer) schools could reach these thresholds when compared to a baseline cohort and thus earn different SPF ratings.

Finally, an important caveat when interpreting both baseline and cohort MGPs is that a statewide (or schoolwide) MGP greater than 50 in a particular grade and year does not guarantee that the overall level of achievement is higher than it was in previous years. At the statewide level, for example, if this year's 4th grade cohort had lower 3rd grade prior scores than the baseline cohort, then this year's 4th graders could make more progress (conditional on their 3rd grade scores) and yet still have lower average 4th grade scores than the baseline cohort. The reverse could happen as well – if this year's 4th grade cohort had earned particularly high 3rd grade scores, and the statewide average baseline SGP was less than 50, it could still be the case that average 4th grade scores were higher this year than they were in the baseline cohort. The same is true at the school level – higher or lower MGPs, whether based on a cohort or baseline model, do not necessarily imply students have reached higher or lower levels of achievement. This is a further reason MGPs, whether based on a cohort or baseline reference, should be interpreted alongside additional data such as average test scores.

Practical Considerations with Baseline SGPs

The use of baseline SGPs also introduces additional practical challenges and assumptions. First, baseline SGPs require that comparable tests be used across all years included in the calculations. To compare 2019 4th graders (who have 2018 3rd grade scores) to 2018 4th graders (who have 2017 3rd grade scores) assumes the same tests were used across years. The 3rd and 4th grade tests do not need to be identical or to be on a vertical scale, but scores on the 2019 and 2018 4th grade tests need to be directly comparable, as do scores on the 2017 and 2018 3rd grade tests. In the cohort SGP model, comparability of the 3rd and 4th grade tests from one year to the next is not required. The baseline SGP model thus places additional pressure on the cross-year equating processes used for the assessments, as the model makes stronger assumptions about score comparability across years.

Second, the use of baseline-referenced SGPs requires a minimum of three years of stable assessment data – the first two years of data can be used to set the baseline, and data from the third year can be compared to the baseline. With only three years of data, however, the model would only be able to incorporate a single prior year scale score. Like many other states, Colorado currently uses multiple prior year scale scores (when they are available) to estimate growth. If the state wanted to continue the use of multiple prior year scale scores, this would require additional years of a stable testing program before baseline SGPs could be computed and interpreted. In order to compute baseline-referenced SGPs that use two prior year scale scores, for example, four total years of data would be required: three years of data would be required to set the baseline and then SGPs could be compared to the baseline in the fourth year. Experts including the developers of the SGP methodology (Betebenner et al., 2014) recommend having at least four years of a stable testing program before using baseline-referenced SGPs operationally.

Third, even if the tests remain stable for a sufficient number of years, baseline SGPs may be sensitive to other changes in test-taking policies. Changes to Colorado's test-taking policies for 8th grade mathematics provide an example. In 2018 8th-graders had the option to take a subject-specific mathematics test instead of the general grade-level test and these students' scores were not used in 8th-grade SGP calculations.[3] In 2019, however, all 8th-graders were required to take a single grade-level 8th-grade mathematics test. In a baseline-referenced SGP model, the performance of all 8th graders in 2019 would be compared to the performance of previous 8th graders who had similar prior year scores and did not elect to take subject-specific tests. If students electing to take subject-specific tests in prior years differ systematically from the population of all 8th graders, the baseline cohort may no longer provide an appropriate or optimal reference group even though the 8th-grade mathematics test did not change.

Finally, it is worth pointing out an interpretational tradeoff when using baseline SGPs. Consider again the earlier hypothetical example in which all districts adopt a new curriculum, and suppose further that all relevant tests have remained the same. Should we compare current 4th-grade students' scores to that of prior cohorts who experienced a fundamentally different 4th-grade instructional experience? While the baseline SGP model would potentially provide a way to evaluate whether there is evidence that the new curriculum was more effective than the prior curriculum (as evidenced by higher growth rates), it may not provide the most useful reference group. Depending on the intended uses and interpretations of the SGPs, it may be more appropriate to compare each student's performance to their own cohort of students, who have had a shared educational experience.

As this discussion hopefully makes clear, choosing between baseline and cohort SGP models requires careful deliberation. In addition to understanding the different data requirements and interpretations for cohort and baseline-referenced SGPs, these deliberations should also consider the practical impact that switching to baseline-referenced SGPs could have. To better understand these implications, the next section provides empirical evidence comparing cohort-referenced and baseline-referenced SGPs and MGPs for 2019 Colorado schools and students.

[3]In 2018 approximately 10,000 out of 60,000 8th-graders took subject-specific mathematics tests.

Empirical Analyses

This section presents empirical comparisons of cohort-referenced and baseline-referenced SGPs based on Colorado test score data from 2015-2019. These analyses are intended to answer two primary questions:

1. How different would inferences about 2019 student and school-level growth be if they were computed using the baseline-referenced SGP model instead of cohort-referenced SGP model?

2. Are there systematic patterns in the differences, and if so, what can they tell us about the use of baseline-referenced versus cohort-referenced SGPs?

This analysis focuses on Mathematics and English Language Arts (ELA) test score data, because these assessments are administered to every student in grades 3-8 and are the primary assessment data used in the School and District Performance Frameworks. Colorado transitioned to the Colorado Measures of Academic Success (CMAS) assessment system in 2014 for Social Studies and Science (Colorado Department of Education, 2014) and in 2015 for Mathematics and ELA (Colorado Department of Education, 2018). From 2015-2017 the Mathematics and ELA CMAS tests were directly based on the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium tests. In 2018 the tests were revised to be shorter, but were still constructed to yield scores that were comparable to the PARCC tests administered in 2015-2017 (Colorado Department of Education, 2018). Although 2019 was the fifth consecutive year of the PARCC/CMAS tests, low participation rates and other technical challenges due to the test transition in 2015 resulted in four years of complete and reliable data, from 2016-2019. Some 2015 scores were used as prior year scores when available. As described above, calculating baseline-reference SGPs with multiple prior year scores can only be done in the 4th (or later) year of a testing program. As a result, the currently available data only allow baseline-referenced to cohort-referenced SGP comparisons for the spring 2019 data. We describe some additional relevant changes that occurred within the CMAS testing program from 2015-19 further below.

Data

In this section we provide comparisons of cohort and baseline ELA and Mathematics SGPs for students enrolled in grades 4 through 8 in 2019, computed in the following two ways:

- **Cohort SGP**: using the 2019 cohort of students, we calculate SGPs that condition on up to 2 prior year scale scores. More specifically, SGPs are computed for each student with a valid 2019 scale score who also has a valid prior year (2018) scale score from the previous grade. If a student has two prior year scale scores (i.e., from 2017 and 2018), these are both used. These SGPs use the standard cohort-referenced model, in which the current cohort of students is the reference group for each grade.

- **Baseline SGP**: we also compute SGPs for the same sample of 2019 students using a baseline cohort as the reference group. The baseline cohort comprises all students who had both a valid current and prior year test score in 2016, 2017, or 2018. Again, when students had two valid prior year scores, these were also used to construct the model. The earliest prior year scores used were from 2015, so that students with 2016 scores only ever had

one prior year score, while students in 2017 and 2018 could have had one or two prior year scores. This combined 2016-2018 super cohort serves as the "baseline cohort" for calculating the 2019 "baseline-referenced" SGPs.

As noted above, there were at least three relevant changes to the CMAS testing program during this time span that need to be considered when interpreting the baseline-referenced SGPs. These changes include:

1. Beginning in 2018 the CMAS test was shortened relative to the CMAS/PARCC test administered in 2015-2017 (2018 Tech Report). Psychometric analyses were carried out to equate scores on the 2018 and 2019 versions of the CMAS with earlier versions, but this still represents a substantial change to the overall test-taking experience for students.

2. From 2015-2017 scale scores are computed using an inverse test characteristic curve (TCC) approach (Thissen & Wainer, 2001), so that students' scale scores are based on their raw scores (Pearson, 2018). Beginning in 2018 (Colorado Department of Education, 2018), a different technique known as "pattern scoring" was used to compute scale scores. In pattern scoring, the entire pattern of a student's responses, not only the number of correct answers, is used to calculate a scale score. As reported in Appendix Tables A1 and A2, the number of unique observed scale scores was substantially lower in some grades in 2016 and 2017 relative to 2015, 2018, and 2019, likely due to this change in the scoring approach. Because SGPs are based on the observed scale scores, these changes could potentially undermine the comparability needed to appropriately interpret baseline-referenced SGPs.

3. Beginning in 2019, the mathematics test-taking patterns for 7th and 8th grade students changed. In 2015-2018, students in 7th and 8th grade could either take the grade-level PARCC/CMAS test or take subject specific math tests (Integrated Math I/II/III, Geometry, or Algebra I/II; https://www.cde.state.co.us/assessment/newassess-parcc-archive). In 2019, all 7th and 8th grade students were required to take a single CMAS grade-level math test. Note that in 2015-2018, the subject mathematics tests are not included in SGP computations.

Tables A1 and A2 in the Appendix present sample sizes and means and standard deviations of scale scores for all students with valid scores in 2015-2019 for ELA and Math, respectively. The tables also report the mean and standard deviation of scale scores for students with valid SGPs in 2017-2019 separately, which are used in the analyses below. For example in 2019, across grades and subjects the SGP sample includes approximately 90 to 95% of all students with valid scale scores. Tables A1 and A2 also report the number of unique scale score values observed in each year and grade. The PARCC/CMAS score scale includes integer values from 650-850, so that there can be up to 201 unique scale scores in theory. In both ELA and Math, there was a noticeable drop in the number of unique observed scale scores in 2016 and 2017 relative to 2015, and then an increase in the number of unique scores in 2018 and 2019. The changes were more pronounced in Math where, for example, in 6th grade there were 193 unique score values observed in 2015, then only 154 and 158 in 2016 and 2017 (respectively), and then 200 in 2017 and 2018. While these changes alone do not undermine the interpretation of baseline-referenced SGPs, such large shifts in the distribution of observed score values provide suggestive evidence that there might be relevant changes to the score scale properties during these years that would need to be further investigated. We discuss this issue below.

The SGPs for these analyses were produced for the Colorado Department of Education in April of 2020 by the National Center for the Improvement of Educational Assessment. Baseline-referenced SGPs were produced by first estimating the necessary coefficient matrices using the 2016-2018 cohorts (subject to data restrictions described above), and then using these coefficients to compute SGPs for students taking tests in 2019. The cohort-referenced SGPs used in these analyses differ slightly from those produced for operational accountability reporting. While operational SGPs use all available prior test scores, the cohort and baseline SGPs in these analyses use at most 2 prior year scale scores because it was not possible to construct the necessary baseline coefficient matrices with additional prior year scores. While it would have been possible to use additional prior year scores to construct the 2019 cohort SGPs, we limited these to 2 prior year scores to keep the samples consistent when comparing across methods. As a sensitivity analyses, the analyses were also conducted using only a single prior year scale score, and results were similar. We focus on the results using up to two prior year scale scores in these analyses because that is closest to what would be used operationally.

Methods

We begin by reporting descriptive statistics for statewide mean scale scores over time by grade and subject, as these can provide initial insight into differences we might expect to see in the SGP data. As noted above, for example, if 4th grade average scores were consistently trending up across years while students' 3rd grade average scores remained similar, it would suggest that 4th grade students were making more progress in more recent years. This trend would be expected to show up in baseline SGPs, but not cohort SGPs.

We next provide summary descriptive statistics that compare the baseline and cohort SGPs at the student level. We provide the statewide mean, median, and standard deviation of each SGP type by grade and subject. We also report the root mean squared difference (RMSD) between cohort and baseline SGPs across students, the correlation between cohort and baseline SGPs across students, and the correlation between each student's baseline SGP and the student's prior year scale score. The RMSD quantifies the average magnitude of the difference between cohort and baseline SGPs for each student. An RMSD is similar to a standard deviation – for example, an RMSD of 3 would indicate that each student's baseline-referenced SGP differs from the student's cohort SGP by 3 points, on average. The correlation between cohort and baseline SGPs indicates how similarly the two SGP types would rank order students. The Pearson correlation coefficients, which can range from -1 to +1, quantify the degree of association between the different SGP types and prior achievement of students. Correlation values near +/- 1.0 indicate strong associations. If students were to be rank ordered on two different metrics with a correlation near 1.0 the rank orderings would be nearly identical. Correlations near 0 indicate no linear associations and suggest that the rank ordering across the two measures could differ substantially. The correlation between baseline SGPs and prior year scale scores can be used to better understand any systematic differences between the two SGP types. By construction, the correlation between cohort-referenced SGPs and prior year scale scores is exactly 0, but this need not necessarily be the case for baseline-referenced SGPs.

We then summarize differences in median SGPs (MGPs) at two levels: the school-by-grade and school-by-EM (elementary/middle) levels. Comparing the differences at the aggregate MGP level is more policy relevant, as high-stakes accountability decisions are based on aggregate MGPs,

not individual student-level SGPs. The school-by-EM level analyses use a single MGP, pooled across grades 4-5 (E) or 6-8 (M). Most schools either enroll students in grades K-5 or 6-8, so that school-by-EM MGPs are essentially school-level MGPs. This is the level of aggregation used for school accountability purposes in Colorado.[4] We also compute MGPs at the school-by-grade level. Although school-by-grade MGPs are not used directly in state accountability reporting, the results may provide diagnostic insights into differences between cohort and baseline SGP and MGP values. MGPs are calculated only for school-grade or school-EM groups with at least 5 observed SGPs. While the state uses a threshold of only computing MGPs when there are at least 20 valid SGPs, we use the lower threshold to describe more of the data.

We first describe differences in aggregate MGPs using the average difference between MGPs and the RMSD between baseline and cohort MGPs. The RMSD again indicates the magnitude of the average difference between MGP types across schools; however, the RMSD does not indicate whether baseline MGPs tend to be higher or lower than cohort MGPs. The average difference in the two types of MGPs tells us whether, on average, baseline-referenced MGPs are higher or lower than cohort-referenced MGPs. A positive average difference would indicate that baseline MGPs are higher, which would suggest student growth is increasing over time, while a negative average difference would indicate baseline MGPs are lower, on average, which would indicate student growth is decreasing over time. According to the accountability theory of action, we hope to see the former pattern (increasing growth over time). We then compute correlations between the two different MGP types, between each MGP type and school-level average prior achievement, and between the difference in MGP types and average prior achievement. Again, the correlation between baseline and cohort MGPs indicates how consistently the two different metrics would rank order schools, while the correlations with prior achievement can be used to help understand whether any differences between MGP types appear to be systematically related to students' prior achievement.

Results

Average Scale Scores.

Figure 1 shows trends in average scale scores across the years 2015-2019, separately for each grade and subject. The averages shown are for all students with a valid scale score. Tables A1 and A2 in the Appendix report the sample sizes, means (shown in Figure 1), and standard deviations of scale scores for all students with valid scale scores. Figure 1 shows that average ELA scores do appear to be trending upwards over time across cohorts. For example, 3rd graders in 2019 had higher average scores (740) than 3rd graders in 2015 (735), which is approximately a 0.10 increase in standard deviation units – a small but nontrivial change in the distribution. Similar changes were observed at other grade levels. The CMAS scores are not vertically linked across grades, so means cannot be compared directly across grades. Although the upward trends may intuitively suggest that growth in achievement has been increasing over time, this is not necessarily true. The upward trend in scores in 3rd grade across cohorts suggests that students are finishing 3rd grade with higher levels of achievement in more recent years. This in turn suggests that while the higher average scores in other grades may represent evidence of faster progress from grades 3-8 in more recent years, it could also reflect students

[4]Here we define the E/M levels based only on grade level and not on the formal CDE designations – so all 4th and 5th grade scores within a school are treated as an "E" unit, while scores from grades 6-8 would be an "M" unit. The official E/M units defined for accountability occasionally include exceptions to these classifications (e.g., some "E" units may include grade 6 data).

beginning 3rd grade at more advanced levels and then maintaining high performance through to 8th grade without making more per-grade progress in these later grades. The changes could also be due to demographic or other shifts in the student population over time.

Trends in Math scores are more mixed – the trends mostly appear flat or slightly negative (6th grade Math scores) with the exception of 8th grade. Although scores cannot be compared across grade levels directly, the contrasting trends in 5th and 6th grade are counter-intuitive, as it suggests that although students in more recent years finished 5th grade with higher test scores, on average, students in more recent years tended to finish 6th grade with lower test scores, on average. The substantial change in 8th grade scores is likely due to changes in the population of students taking the 8th grade CMAS test described above, and should not be directly interpreted as a change in the average achievement of students statewide. Table A2 indicates that in 2015-2017 there were valid scale scores for between 40,562 and 43,158 students; there were 49,189 scores in 2018 and 58,863 in 2019. These changes suggest that the additional students taking the 8th grade math test in more recent years, who had previously taken the subject-specific tests, tended to be higher-achieving students on average. The apparent trend in 8th-grade scores is more appropriately attributed to a changing population of students taking the tests.

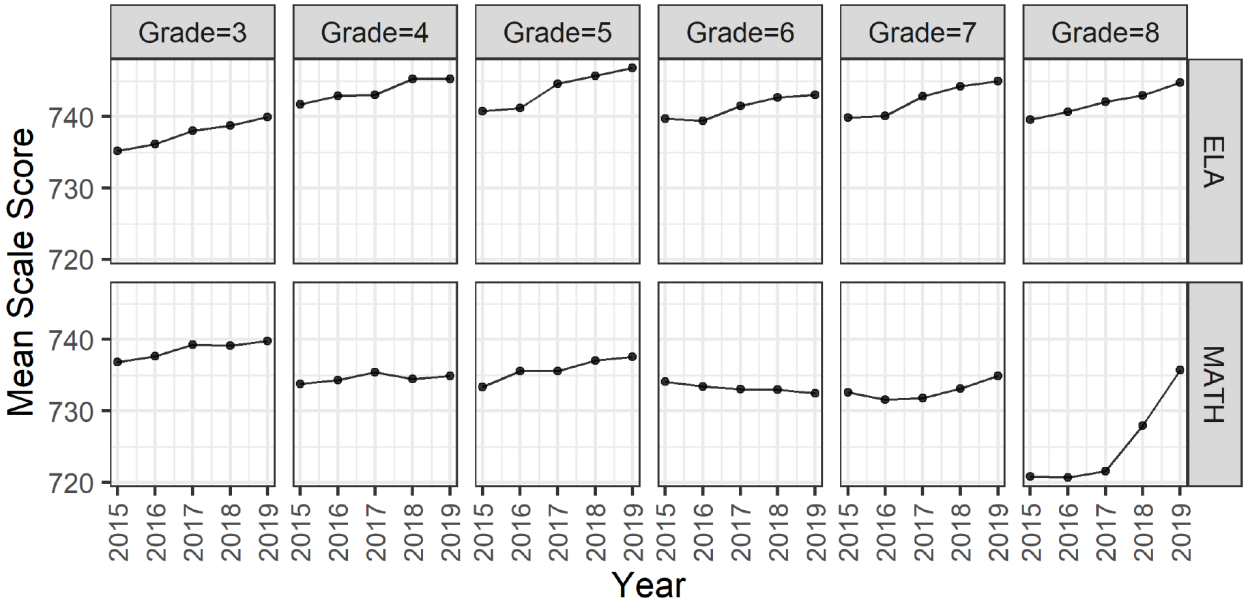

Figure 1. Average Scale Scores by Grade, Subject, and Year.

The trends in average 7th and 8th-grade Math scores need to be interpreted differently than in other subjects due to changes to the test-taking population and policies described earlier.

Student-Level SGPs.

Table 1 summarizes the distributions of 2019 cohort and baseline SGPs across students by subject and grade. Within each subject and grade there are SGPs for between approximately 53,000 and 62,000 students. As expected, the statewide mean and median cohort SGP are almost exactly 50 across all grades, with a standard deviation of approximately 29 points. The mean and median baseline SGPs are more variable across grades. In ELA, the mean and median baseline SGPs are always equal to or less than 50. In Math, the mean and median baseline SGPs are sometimes higher than 50 and sometimes lower and are generally more variable than in ELA. Taken at face value, these results suggest that despite the upward trend in ELA scores across cohorts, at each grade level students in the most recent cohort are making about the same amount of progress or even slightly less than students in the baseline cohort. In Math the results suggest that in 5th, 7th, and 8th grade, 2019 students made more progress than prior cohorts, whereas in 4th and 6th grade, students made less progress. Figure 2 displays histograms of the differences between cohort and baseline SGPs across students. The histograms suggest that although the average differences are generally not too far from 0 (as expected based on the values in Table 2), there are some grades in which the differences appear to be systematically positive or negative. For example, in 6th grade math the majority of students had higher cohort SGPs than baseline SGPs, while in 7th grade math the opposite is true.

Table 1. Summary Statistics for 2019 Cohort and Baseline SGPs, by Subject and Grade.

			Cohort SGP			Baseline SGP				Cor(Base, Prior)
Subject	Grade	N	Mean	Median	SD	Mean	Median	SD	RMSD	
ELA	4	58726	50.0	50	28.9	49.7	50	28.9	2.8	0.02
ELA	5	60687	50.1	50	28.9	50.0	50	28.7	2.1	-0.02
ELA	6	60091	49.9	50	28.9	48.9	48	29.3	2.7	-0.03
ELA	7	58033	50.0	50	28.9	47.7	47	29.1	4.3	-0.09
ELA	8	54081	49.9	50	28.9	49.4	49	29.9	2.6	-0.02
Math	4	60316	50.0	50	28.9	48.4	48	28.8	3.4	-0.06
Math	5	61823	49.9	50	28.9	51.2	52	29.2	3.5	-0.01
Math	6	60216	50.0	50	28.9	46.0	44	28.6	6.3	-0.13
Math	7	58087	50.0	50	28.9	52.7	54	29.2	3.7	0.02
Math	8	53159	49.9	50	28.9	51.9	53	29.8	3.1	0.04

For any individual student, however, the differences are very modest. The correlation between cohort and baseline SGPs across students within grades is 0.99 or higher in every grade (values not shown in Table 1). This suggests that students with higher cohort SGPs also tend to have higher baseline SGPs. The RMSD is between about 2 and 6 points in each grade, suggesting that on average each student's SGP would differ by between 2 and 6 points depending upon which method was used. These differences are far smaller than the uncertainty in individual student SGPs; the average standard error of each student's cohort SGP is approximately 14-18

points, depending on the grade and subject. Thus, at the student level, SGP differences smaller than about 15 points are generally not meaningfully different from chance variability; although for students with particularly high or low SGPs a change of 6 points could be larger than expected due to chance, it would not substantively alter the interpretation of the student's SGP. These results suggest that cohort versus baseline SGP differences at the individual student level are generally not large enough to be meaningful.

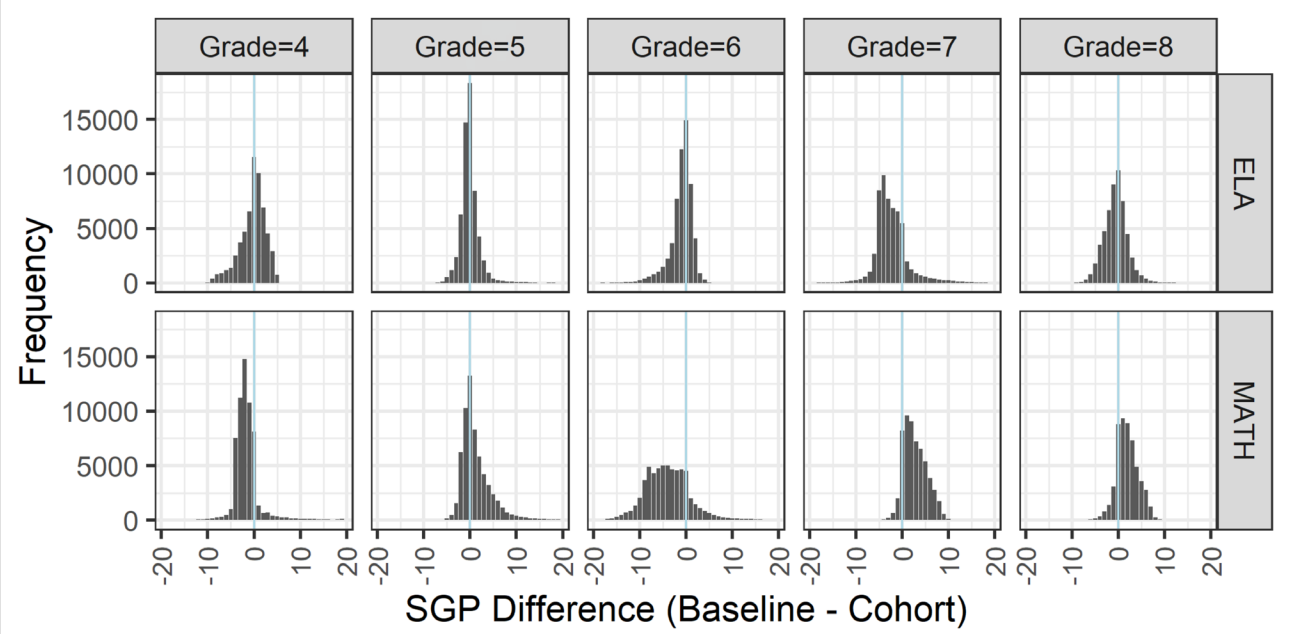

Figure 2. Histogram of Baseline-Cohort SGP Differences, by Subject and Grade.

Blue vertical lines indicate X=0. Differences are computed by subtracting cohort SGPs from baseline SGPs. A total of N=136 students (across all grades and subjects) with absolute differences greater than 20 are excluded from the histograms to facilitate more easily interpreted plots.

The final column in Table 1 shows the correlation between students' baseline SGP and their prior year test score. These correlations are generally small but not exactly 0 – they range from -0.13 to 0.04. In contrast, the correlation between each students' cohort SGP and their prior year scale score is exactly 0 by construction. The correlations suggest that, on average, students with higher prior year test scores tended to have slightly lower baseline SGPs. This pattern prompted us to look more closely at the relationship between prior year scale scores and baseline SGPs. Figures 3 and 4 show the association between the difference in each student's cohort and baseline SGPs and their prior year scale scores, separately by grade and subject. Figure 3 shows the difference between each student's baseline and cohort SGP on the y-axis relative to their prior year scale score on the x-axis. Figure 4 shows the median baseline SGP on the y-axis across 10 equally spaced bins of prior year scale scores shown on the x-axis. Because the median cohort SGP is 50 at every prior year scale score value by construction, the two figures reflect similar patterns; when students tend to have higher baseline SGPs in Figure 3 (indicated by positive differences), the median baseline SGP will tend to be greater than 50 in Figure 4.

Figures 3 and 4 suggest that differences between a student's cohort and baseline SGP are

systematically related to prior year scale scores in some grades, particularly in 4th and 6th grade Math. In 4th grade Math (the lower left panel of both figures), for example, nearly every student with a prior year (2018 3rd grade) scale score above 700 had a lower baseline SGP than cohort SGP; as a result, the median baseline SGPs were consistently below 50 for these students. If a student has a lower baseline SGP than cohort SGP, we would infer that the student made less progress when compared to the baseline cohort than when compared to other students in their own cohort. Mathematically, this occurs because at a certain level of prior achievement, students in the current cohort earned lower current year test scores than students with similar prior achievement in the baseline cohort. Why might this happen? There are at least two potential explanations, and there could potentially be others.

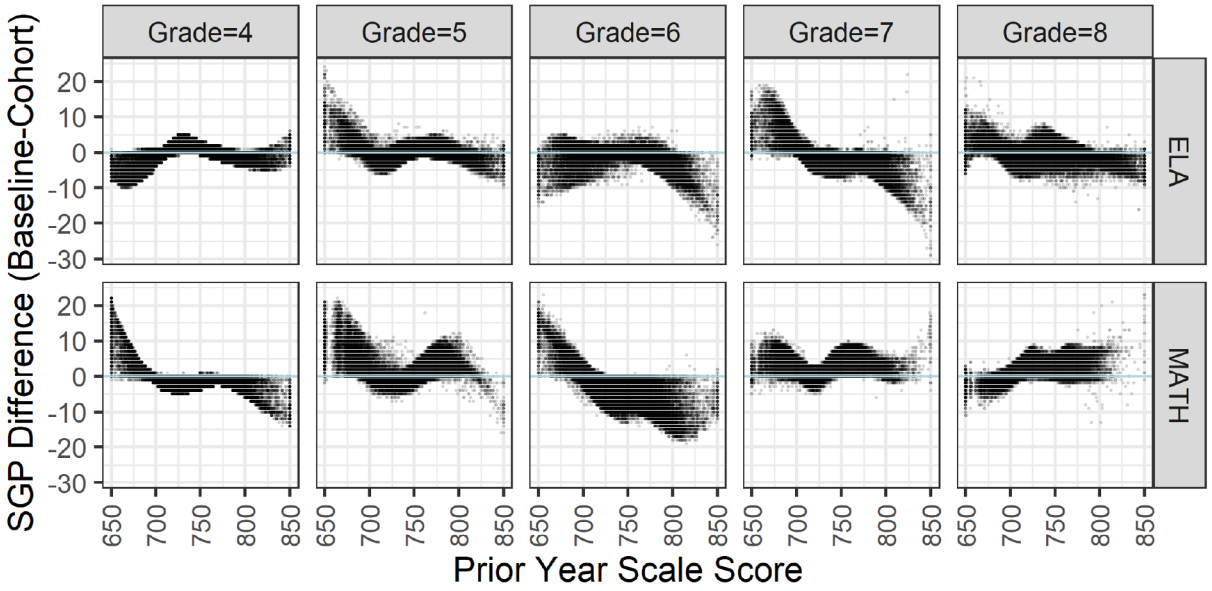

Figure 3. Difference Between Baseline and Cohort SGPs versus Prior Year Scale Score, by Grade and Subject.

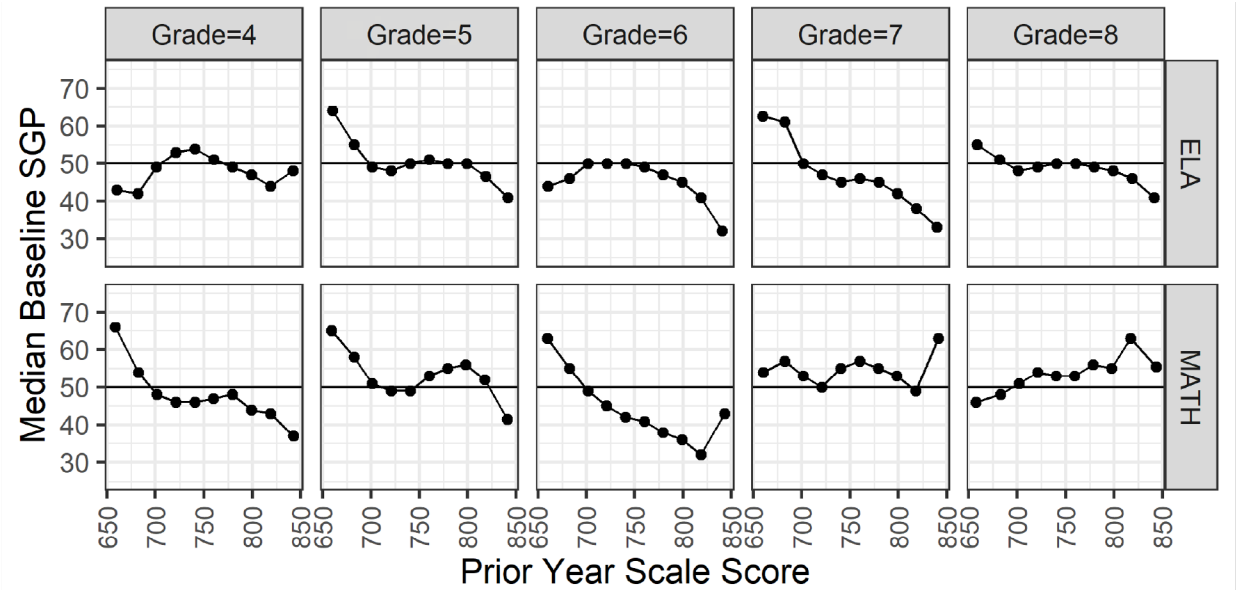

Figure 4. Median Baseline SGP versus Average Prior Year Scale Score, by Grade and Subject.

First, these differences may reflect real differences in student learning. It may be that, conditional on prior achievement, students in the current cohort did reach higher or lower levels of achievement relative to content tested in the current grade. In the case of 4th grade Math here, we would conclude that higher-achieving students in the current cohort made less progress than similar students in the baseline cohort. Second, the differences could result from uncertainty in test equating rather than differences in learning. Slightly different test forms are used each year, and adjustments are made so that scores can be compared across years using a method known as test equating. Because the baseline SGPs compare students to prior cohorts, uncertainties in the test equating process can, in theory, carry over to the resulting baseline SGPs. This does not arise with the cohort-referenced SGP model, because students are only compared to other students in their own cohort, who took the same grade-level tests in the same years. While these results cannot determine the cause of the differences, it is important to note that uncertainty in the equating process is a potentially relevant factor when interpreting baseline SGPs but not when interpreting cohort SGPs.

These comparisons suggest important avenues for further investigation. First, are there plausible explanations as to why students with higher or lower prior year achievement scores might have made more or less progress in 2019, and why these patterns might vary across grades and subjects? Can these explanations account for the sharp discontinuities observed in Figure 3? Second, and related, what patterns of differences between cohort and baseline SGPs would be expected? What other changes in learning might we expect to see and what would those look like in Figures 3 and 4? For example, according to the accountability theory of action, we might anticipate that students in schools with lower prior year scores, which were provided additional supports, to make more progress in the current year. However, it is less clear why there would also be a decrease in SGPs for students with higher prior scores. Finally, what effects might uncertainty in test equating or different types of equating designs have on baseline SGP calculations?

Aggregate School-Level MGPs.

Patterns in differences between MGPs at the school by grade and school by Elementary/ Middle (E/M) level follow similar patterns to those seen in the student-level results. Table 2 reports the average differences in cohort and baseline MGPs, as well as the RMSD and correlation between the different types of MGPs, by subject, grade and E/M level. Table A3 reports summary statistics for these MGPs in 2019. As with the student-level data, the average differences between cohort and baseline MGPs were negative, although generally small, across all grades and E/M levels for ELA. In Math the differences are negative (but small) at the E/M level, but at the grade level, the average differences were sometimes negative and sometimes positive depending on the grade in Math. This suggests that there is no correlational evidence of systematic increases in student learning in the 2019 cohort of students, relative to the 2016-2018 cohorts. The RMSDs ranged from about 1.6 to about 6.2 points depending on the grade and subject. The RMSD indicates the average cohort versus baseline-referenced MGP across schools, suggesting that across grades and subjects, school MGPs differed by at most 6 points on average. At the E/M level, the RMSDs ranged from 1.5 to 3.5 points, again with larger differences in Math than in ELA.

To evaluate the practical magnitude of the RMSDs, we compare them to two different quantities. First, the standard deviation of cohort MGPs across schools in 2019 ranged from approximately 11 to 16 points, depending upon the grade and subject. Relative to this between-school variability, differences of about 1-3 points do not seem particularly large, but differences of 6 points could be considered moderate. Second, we looked at the change in cohort MGPs for each school by grade or school by E/M level from 2017 to 2018 and then 2018 to 2019. In other words, we calculated the amount we might expect a school's (cohort) MGP to change from one year to the next. There are many reasons a school's MGP would be expected to change from one year to the next in addition to uncertainty in the MGPs – for example, schools may make instructional or curricular changes and new cohorts of students enter the school. These differences are summarized in Table A4 in the Appendix. The RMSDs were approximately 15-16 points on average across grades and subjects at the school by grade level, and approximately 13 points at the E/M level. Thus, the differences between cohort and baseline MGPs within 2019 tended to be only about 10-50% as large as the differences observed in cohort MGPs from 2018-2019 or 2017-2018. The cross-year correlations between cohort-referenced MGPs ranged from as low as 0.28 (8th grade ELA from 2017 to 2018) to 0.62 (Middle school math from 2017-2018), and were approximately 0.45 on average across grades, subjects, levels, and years. Relative to the cross-year differences, the differences between cohort and baseline MGPs within 2019 are generally small.

Table 2. Mean Differences, RMSDs, and Correlations between Baseline and Cohort Median Growth Percentiles by Subject, Grade, and E/M Levels.

Subject	Level	N	Mean Diff.	RMSD	Cor 1	Cor 2	Cor 3	Cor 4
ELA	4	1050	-0.22	2.00	0.99	0.22	0.24	0.15
	5	1059	-0.11	1.59	0.99	0.12	0.10	-0.21
	6	639	-1.29	2.10	1.00	0.22	0.19	-0.24
	7	560	-2.85	3.73	0.99	0.20	0.09	-0.62
	8	559	-0.44	1.84	0.99	0.13	0.11	-0.10
Math	4	1050	-2.11	2.90	0.99	0.14	0.08	-0.52
	5	1059	1.52	2.70	0.99	0.04	0.04	0.04
	6	638	-5.24	6.17	0.98	0.14	-0.01	-0.70
	7	560	3.77	4.14	0.99	0.22	0.26	0.36
	8	559	2.42	3.00	0.99	0.19	0.23	0.48
ELA	E	1090	-0.13	1.54	0.99	0.31	0.31	-0.05
	M	731	-1.52	2.20	0.99	0.32	0.28	-0.32
Math	E	1091	-0.19	2.05	0.99	0.24	0.18	-0.31
	M	730	-0.70	3.49	0.97	0.28	0.23	-0.24

Note: Cor 1 = corr(cohort SGP, baseline SGP); Cor 2 = corr(cohort SGP, prior scale score); Cor 3 = corr(baseline SGP, prior scale score); Cor 4 = corr(MGP difference, prior scale score).

Finally, Table 2 presents correlations between different MGP and prior score metrics. The correlations between the 2019 cohort and baseline MGPs ("Cor 1" in Table 2) were 0.97 or higher across all grades and E/M levels, suggesting that the rank ordering of schools would remain very consistent regardless of which SGP type were used. In addition, these correlations were substantially higher than the between-year correlations in MGPs reported in Table A4. The high correlations and relatively small magnitude of the average differences and RMSDs between cohort and baseline MGPs suggest using baseline SGPs in place of cohort SGPs would not be anticipated to have large effects on school SPF ratings. However, because the SPF calculations are based on MGPs also for demographic subgroups and are relative to fixed thresholds, more analysis would be needed to determine the actual impact the change would have on SPF ratings.

Columns "Cor 2" and "Cor 3" in Table 2 report the correlation between each school's cohort (Cor 2) or baseline (Cor 3) MGP and students' average prior year scale scores. Prior studies have shown that measurement error in student test scores can cause bias in cohort MGPs that would tend to favor schools enrolling students with higher prior test scores (McCaffrey et al., 2015; Shang et al., 2015), an issue we investigate further in an accompanying report. The correlations in Table 2 do suggest that, on average, schools with higher cohort or baseline MGPs tend to be schools where students had slightly higher prior year average test scores, although the correlations are generally small. More relevant to the present analyses, these correlations appear to be similar for both the cohort and baseline SGPs. The primary exception is 6th grade Math, where the correlation between baseline MGPs and average prior achievement is actually slightly negative; although this is the only negative correlation, it is so close to 0 that it does not represent a meaningful negative association. In addition, the differences in baseline and cohort SGPs for 6th grade Math have already been identified as needing further analysis.

The final column of Table 2 ("Cor 4") shows the correlation between the baseline/cohort MGP difference at each school and students' average prior year scale scores. These correlations suggest strong associations between the prior achievement of students at a school and the difference between the cohort and baseline MGPs at the school for some grades and subjects. But these correlations are also highly variable and do not indicate a clear pattern. In 8th grade Math, for example, the correlation is 0.48, suggesting that schools where 8th graders had higher prior year scores tended to have higher baseline MGPs relative to cohort MGPs. On the other hand, the correlation of -0.70 for 6th grade Math suggests that schools enrolling 6th graders with higher prior year scale scores tended to have lower baseline MGPs relative to cohort MGPs. At the E/M level, the correlations are all negative, suggesting that when aggregated to the E/M level schools enrolling students with higher prior achievement tended to have lower baseline MGPs relative to cohort MGPs. Figure 5 shows the difference between cohort and baseline school MGPs versus average prior year scale scores by subject and grade. The scatterplots emphasize the highly variable nature of the association between the difference in MGPs and average prior year scale scores. The plots illustrate that 6th grade Math again appears unusual relative to the patterns observed in other grades and subjects. There is no clear explanation for this variability, and the same questions posed earlier about differences in student SGPs would apply to these school-level results in terms of identifying anticipated patterns or trying to identify explanations for the observed variability.

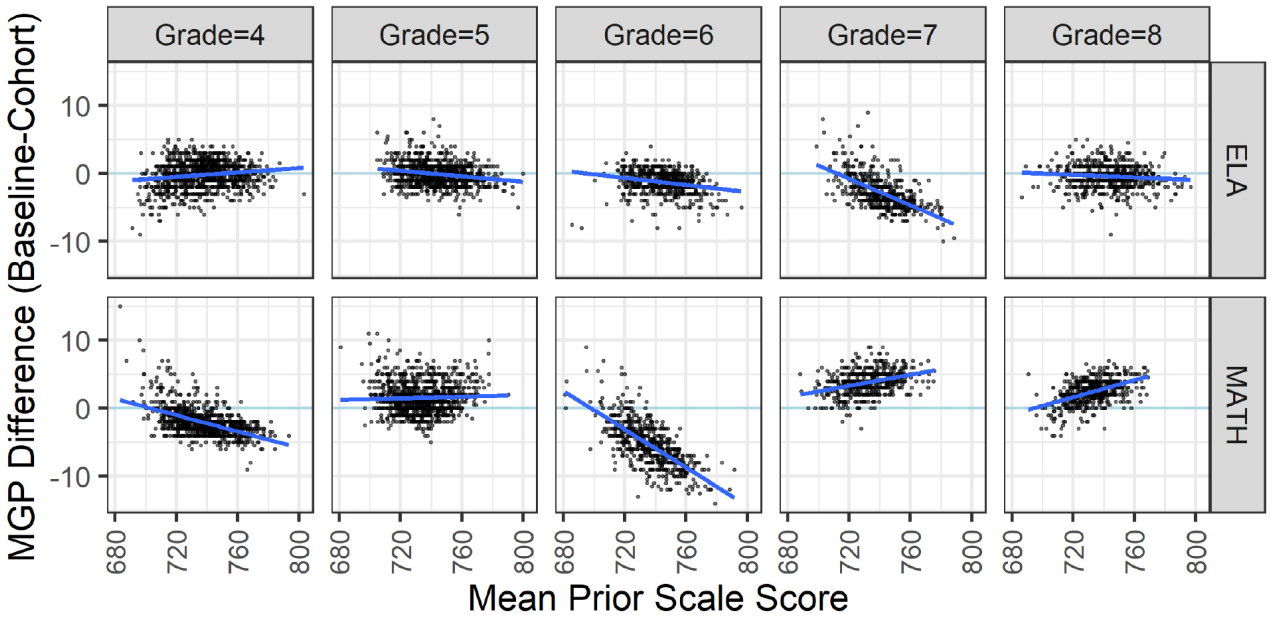

Figure 5. Scatterplots of Difference between School Cohort and Baseline Median Growth Percentiles and Mean Prior Year Scale Scores, by Subject and Grade.

Summary

Colorado reports SGPs to quantify each student's growth. A student's SGP compares their current test score to other students with similar prior scores, and is intended to provide an additional perspective on their performance beyond the current year test score alone. At the school level, MGPs describe the relative progress of students enrolled in each school. Because SGPs are norm-referenced, SGPs and associated MGPs can vary depending upon the norm group used to compute them. Colorado uses the current cohort of students as a norm group, and this cohort-based approach results in an overall average and median SGP of 50 in each grade each year. As a result, SGPs computed in this way cannot be used to answer questions about whether students in more recent cohorts are making more or less academic progress each year and they are sometimes viewed as imposing a "zero sum game" in which some schools and students will always have below average MGPs in a given year (Betebenner et al., 2014).

Baseline-referenced SGPs provide a method that can apparently overcome this. By comparing student performance to a stable "baseline" cohort that does not change over time, it is possible in theory to answer questions about whether students in more recent cohorts are making more or less progress than students in the baseline cohort(s) and there is no restriction on the proportion of students receiving SGPs above or below 50. In practice, however, the use of baseline SGPs places greater requirements on the data that may not always be feasible. For example, because students in the baseline and current cohorts must have taken the same tests, baseline SGPs can only be computed after a stable testing program has been in place for a minimum of four years (Betebenner et al., 2014). In addition, any idiosyncrasies in the linking or equating process used to make scores comparable across years, which will cancel out when students are compared to their own cohort, can be mistaken for changes in student progress when using baseline SGPs. To date there has been little research about how large these effects might be and how best to detect them. Finally, the use of baseline-referenced SGP and MGP statistics does not change the primary nature and interpretation of SGPs and MGPs as descriptive statistics of student progress. While both cohort and baseline-referenced SGPs provide useful descriptive information beyond average test scores, neither type of SGP can provide a direct indicator of school or district effectiveness without additional information. Many factors in addition to students' experiences in school or opportunities to learn can affect their test scores, and these factors are reflected in test scores, regardless of the reference group used.

This report compared cohort SGPs to baseline SGPs for the 2019 cohort of students. While the cohort SGPs were computed using the same methodology used in the SPF, the baseline SGPs compared 2019 students' progress to the progress students made in a "baseline cohort" consisting of students in the 2016-2018 cohorts. There were two main takeaways from comparing these two different types of SGPs both at the student level and at the aggregate school level.

First, the average differences between SGPs and MGPs computed using the two methods were generally small, suggesting that inferences about student growth would be similar whether using the cohort or baseline SGP models. It is possible that over a longer time period there could be

systematic trends in baseline SGPs, but this will only be possible to investigate once there are additional years of a stable testing program. At the student level, the two types of SGPs were nearly perfectly correlated (0.99 or higher in all grades), and the observed differences in SGPs were generally much smaller than the standard errors of each student's SGP. At the school level, average differences between each school's cohort and baseline MGPs were smaller than the observed changes in cohort MGPs across years and the correlations between baseline and cohort MGPs were very high (0.97 and above). The high correlations between cohort and baseline SGPs and MGPs indicate that the rank ordering of students or schools would remain essentially the same regardless of which SGP type were used. Although the high correlations and small average differences in school MGPs suggests accountability ratings are unlikely to differ significantly when using baseline versus cohort SGPs, for schools with MGP values near the SPF thresholds there could be differences in final ratings worth investigating.

The second primary finding was that differences in cohort and baseline SGPs were inconsistent across grades and subjects. This was especially true in Math, where the median statewide baseline SGPs deviated considerably from 50 in some grades, but also showed sharp discontinuities at different points in the prior year score distribution, for example in the 4th and 6th grade results. The differences across grades and subjects illustrate why it can be useful to compare patterns across grades as a diagnostic tool for understanding properties of the tests and growth metrics, but also raise questions about the potential causes of these differences. Two potential avenues for investigation could be whether there were relevant policy or instructional changes implemented that can explain the differences and whether uncertainty in grade-specific test linking and equating could be contributing to these results. Explanations for these patterns should be explored and any key findings from these investigations should accompany results reporting baseline SGPs.

References

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Betebenner, D. W., Diaz-Bilello, E., Marion, S., & Domaleski, C. (2014). *Using student growth percentiles during the assessment transition: Technical, practical and political implications*. Council of Chief State School Officers. https://www.nciea.org/sites/default/files/publications/Using_Student_Growth_Percentiles_During_the_Assessment_Transition_SM14.pdf

Colorado Department of Education. (2014). *Colorado Measures of Academic Success: Science and social studies assessments technical report*. Colorado Department of Education. https://www.cde.state.co.us/assessment/cmas_coalt_techreport

Colorado Department of Education. (2018). *Colorado Measures of Academic Success (CMAS) Mathematics & ELA (including CSLA) technical report*. Colorado Department of Education. https://www.cde.state.co.us/assessment/cmas_coalt_techreport

Data Quality Campaign. (2019). *Growth data: It matters, and it's complicated*. Data Quality Campaign. https://dataqualitycampaign.org/wp-content/uploads/2019/04/DQC-Growth-Data-Resources.pdf

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice, 34*(1), 15–21. https://doi.org/10.1111/emip.12062

Pearson. (2018). *Partnership for Assessment of Readiness for College and Careers (PARCC) final technical report for 2017 administration*. Pearson. https://files.eric.ed.gov/fulltext/ED599198.pdf

Shang, Y., Van Iwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for student growth percentiles using the SIMEX method. *Educational Measurement: Issues and Practice, 34*(1), 4–14. https://doi.org/10.1111/emip.12058

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates.

Appendix

Table A1. Summary of Mean Scale Scores and Sample Sizes for Full and SGP Samples, by Year (ELA).

		All Valid Scores				SGP Sample			
Year	Grade	N	Mean	SD	Unique	N	Mean	SD	Unique
2015	3	62674	735.2	40.0	201				
2015	4	62329	741.7	34.1	201				
2015	5	61954	740.7	32.3	201				
2015	6	60843	739.8	32.0	200				
2015	7	57342	739.9	37.2	201				
2015	8	54533	739.6	37.6	201				
2016	3	63379	736.1	39.4	178				
2016	4	63024	742.9	35.3	184				
2016	5	61980	741.2	32.3	185				
2016	6	60066	739.4	32.1	182				
2016	7	58082	740.1	37.1	189				
2016	8	53904	740.7	37.4	190				
2017	3	63606	738.0	40.0	172				
2017	4	64116	743.0	35.3	185	58882	743.8	35.1	185
2017	5	63391	744.6	34.1	178	58331	745.4	33.9	177
2017	6	60839	741.5	31.3	186	55984	742.0	31.2	185
2017	7	58778	742.8	37.8	185	53623	743.6	37.6	183
2017	8	56211	742.1	38.5	177	50777	743.0	38.3	177
2018	3	63016	738.8	39.8	201				
2018	4	64789	745.3	35.4	201	59469	745.9	35.4	201
2018	5	65359	745.7	34.1	201	59976	746.5	33.9	201
2018	6	63647	742.7	33.5	200	58201	743.1	33.4	200
2018	7	60907	744.2	39.7	201	55423	744.9	39.7	201
2018	8	58684	743.0	39.7	201	52761	743.8	39.7	201
2019	3	60796	739.9	40.7	201				
2019	4	63258	745.3	35.9	201	58726	746.0	35.8	201
2019	5	65757	746.8	34.1	201	60687	747.8	33.9	201
2019	6	64493	743.1	32.8	200	60091	743.5	32.7	200
2019	7	62645	745.0	38.5	201	58033	745.7	38.4	201
2019	8	58808	744.8	40.3	201	54081	745.7	40.2	201

Table A2. Summary of Mean Scale Scores and Sample Sizes for Full and SGP Samples, by Year (Math).

		All Valid Scores				SGP Sample			
Year	Grade	N	Mean	SD	Unique	N	Mean	SD	Unique
2015	3	63766	736.9	34.0	201				
2015	4	62329	733.8	30.5	197				
2015	5	61917	733.3	30.9	197				
2015	6	60749	734.1	30.4	193				
2015	7	55346	732.6	26.6	179				
2015	8	40562	720.8	32.3	196				
2016	3	65013	737.6	35.9	164				
2016	4	63611	734.3	32.3	178				
2016	5	62106	735.6	32.2	167				
2016	6	60346	733.4	31.1	154				
2016	7	55611	731.6	28.0	143				
2016	8	41325	720.7	34.5	163				
2017	3	65420	739.3	37.2	167				
2017	4	65009	735.4	33.0	162	60429	735.8	32.9	162
2017	5	63446	735.6	32.4	165	58837	736.0	32.4	165
2017	6	60950	733.1	32.3	158	56097	733.5	32.2	158
2017	7	56210	731.8	27.4	154	51349	732.2	27.4	151
2017	8	43158	721.6	34.6	165	38723	722.1	34.6	165
2018	3	64714	739.1	36.5	201				
2018	4	65995	734.5	33.2	196	61206	734.8	33.1	196
2018	5	65516	737.0	34.0	201	60818	737.4	33.9	201
2018	6	63765	733.0	31.3	200	58262	733.3	31.3	200
2018	7	59983	733.1	28.7	185	54569	733.5	28.8	185
2018	8	49189	727.9	37.1	201	43294	727.5	36.5	200
2019	3	62560	739.8	36.4	201				
2019	4	64474	734.9	32.5	196	60316	735.3	32.4	196
2019	5	65917	737.6	33.9	199	61823	738.0	33.9	199
2019	6	64650	732.5	31.0	200	60216	732.8	31.0	200
2019	7	62790	734.9	29.4	197	58087	735.4	29.4	196
2019	8	58863	735.8	41.2	201	53159	735.5	40.4	201

Table A3. Summary Statistics for 2019 School MGPs, by Subject and Grade.

Subject	Grade/Level	N	Mean	Median	SD	Min	Max
ELA	4	1050	49.8	49	14.3	8	91
ELA	5	1059	50.2	50	13.1	8	95
ELA	6	639	52.1	51	15.6	10.5	89.5
ELA	7	560	49.9	50	14.4	9	90
ELA	8	559	50.5	50.5	14.3	8	98
ELA	E	1090	49.7	50	11.3	7	92
ELA	M	731	51.4	51	13.7	8	89.5
Math	4	1050	49.6	49	16.2	7	99
Math	5	1059	49.8	49	15.6	6	94
Math	6	638	51.5	52	15.3	4	96
Math	7	560	49.8	49	14.2	8	90
Math	8	559	49.3	49	14.7	5	91
Math	E	1091	49.5	49	13.1	1	93
Math	M	730	51.2	50.5	13.6	5	96

Table A4. Summary of Cross-Year Differences in MGPs by Grade, Subject, and Year.

		2018 to 2019					2017 to 2018				
Subject	Grade	N	Avg. Diff.	MAD	RMSD	Corr.	N	Avg. Diff.	MAD	RMSD	Corr.
ELA	4	1000	0.60	11.51	14.73	0.46	1000	-0.26	12.07	15.64	0.43
	5	999	0.08	11.50	14.72	0.40	999	0.46	12.18	15.72	0.39
	6	570	-0.65	11.33	14.84	0.55	570	0.84	12.29	16.19	0.52
	7	504	-0.55	12.19	15.70	0.40	504	-0.26	13.69	17.47	0.38
	8	504	0.21	13.57	17.53	0.28	504	-0.16	13.08	17.14	0.37
Math	4	999	0.41	12.53	15.95	0.51	999	-0.07	13.05	16.66	0.46
	5	998	0.22	12.80	16.28	0.46	998	-0.37	13.06	16.54	0.47
	6	569	-1.00	10.48	13.50	0.62	569	-0.04	11.57	15.04	0.59
	7	503	-0.62	10.24	13.31	0.50	503	-0.14	10.56	13.68	0.52
	8	489	-0.36	11.98	15.35	0.36	489	-0.30	11.69	15.47	0.38
ELA	E	1042	-0.02	9.30	12.07	0.43	1042	0.48	10.11	13.28	0.39
	M	672	-0.85	10.33	13.84	0.48	672	0.48	11.15	14.77	0.48
Math	E	1042	0.46	10.30	13.33	0.48	1042	-0.44	10.77	13.62	0.49
	M	671	-1.39	8.96	11.70	0.62	671	0.77	9.43	12.52	0.60

Note: MAD=mean absolute difference; RMSD=root mean squared difference; Corr.=correlation.