Maryland State Department of Education K-3 school growth measure exploration

Lisa Dragoset

September 2021

For Maryland State Department of Education

REL MID-ATLANTIC
Regional Educational Laboratory

This memo was funded by the U.S. Department of Education's Institute of Education Sciences (IES) under contract ED-IES-17-C-0006 by Regional Educational Laboratory Mid-Atlantic administered by Mathematica. The content does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

MSDE used PARCC assessments from 2015 through 2019 to measure schools' contributions to student learning, or growth, in reading and in math, annually between grades 3 and 8, using student growth percentiles (SGPs). SGPs examine improvement from wherever students start and provide additional information about a school's contribution beyond proficiency levels.

MSDE began administering its Kindergarten Readiness Assessment (KRA) at the start of the 2014/15 school year. The KRA measures children's knowledge and skills at the point of school entry in four domains: language and literacy, mathematics, social foundations, and physical well-being and motor development. The information on children's development and skills at kindergarten entry can inform decisions about how to support children's learning at the school, district, and state levels. In the 2017/18 school year, this first KRA cohort completed the grade 3 PARCC, which created an opportunity to measure growth from kindergarten entry through the end of grade 3, or K–3, for the first time.

MSDE partnered with Regional Educational Laboratory (REL) Mid-Atlantic to examine whether it was feasible to construct a K–3 SGP growth measure. Equipped with a valid K–3 growth measure, MSDE would be able to identify elementary schools with low- and high-performing early grades, and more effectively guide policy and resources to improve those schools. With more than 28 states using kindergarten entry assessments (Education Commission of the States, 2020), this study provides a blueprint for constructing early elementary growth measures using different assessments that are administered more than one year apart.

The key findings from the study were:

- The overall Kindergarten Readiness Assessment score performs about as well in predicting grade 3 achievement as combinations of kindergarten readiness subscores.
- Schools' K–3 growth estimates are likely less valid than schools' grade 3–4 growth estimates but have a similar level of precision.
- Schools' K–3 growth estimates are much less precise for smaller schools than for larger schools.
- Administering the Kindergarten Readiness Assessment to a subset of students in each classroom (as opposed to all students) greatly reduces the precision of schools' K–3 growth estimates.

To inform the understanding of local school officials in Maryland as well as officials in other states that might consider measuring K–3 student growth, this memo describes the data and methods that MSDE and REL Mid-Atlantic used for this study. In the sections that follow, we first describe the data sources and sample used to calculate the K–3 school growth measure. We then describe how we calculated the growth measure, followed by our approaches for assessing the validity and precision of the measure. We conclude the memo with thoughts about the future of calculating K–3 school growth measures.

B. Data sources and sample

The study used administrative data on assessments and schools attended each year, provided by MSDE. Below we describe the data sources and sample in detail.

1. Data sources

A list of the data files from MSDE is in Exhibit 1. Students, schools, and districts were uniquely identified across these files using student, school, and district identification numbers that MSDE provided. No personal identifiers such as student names were included in the data.

Kindergarten Readiness Assessment (KRA) data. The KRA measures skills and knowledge consistent with Maryland's early learning standards for children of kindergarten age. For example, the KRA measures understanding of common vocabulary; knowledge of shapes, colors, and numbers; and the ability to attend to a

learning task. For each kindergartner administered the KRA in 2014, the data included an overall KRA scaled score; a readiness level based on this overall score (demonstrating, approaching, and emerging); and a scaled score for each of the assessment's four domains: language and literacy, mathematics, social foundations, and physical well-being and motor development. The data file also indicated the school and district where the student took the KRA. (These data excluded scores for students at the Maryland School for the Blind and the Maryland School for the Deaf.) See Exhibit 2 for a description of the KRA assessment.

Partnership for Assessment of Readiness for College and Careers (PARCC) data. For school year 2017/18, for each student, the data included scaled scores for the reading and math subject tests that were administered as part of the spring PARCC in grade 3. The data file also included a proficiency level for each subject based on these scores (ranging from 1 to 5) and indicated the school and district where the student took the spring PARCC. See Exhibit 2 for a description of the PARCC assessment.

Attendance data. For each student and each school year, the data included the grade and school in which the student was enrolled as of the last day of the school year. The study used these data to account for student movement across schools when calculating the school growth measure. In other words, students who transferred schools between the start of kindergarten and the end of grade 3 contributed to the growth measure for each school they attended, based on the amount of time the student was enrolled in each school. The data included the date the student enrolled in each school each year and the number of days in attendance and absent. The study used the attendance data to calculate the amount of time each student was enrolled in each school.

Exhibit 1. Data provided by the Maryland State Department of Education

Data source	How data were used
KRA scores for 2014/15 school year	Kindergarten scores for the K–3 growth measure
PARCC scores for 2017/18 school year	Grade 3 scores for the K–3 growth measure
Attendance data for 2014/15 through 2017/18 school years	Link students to schools and districts

KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.

Exhibit 2. Snapshot of the Kindergarten Readiness Assessment and Partnership for Assessment of Readiness for College and Careers assessment

Assessment	When it is administered	What it measures	How it is scored	How it is scaled
KRA	Kindergarten, fall	Kindergarten readiness across four domains: language and literacy, mathematics, social foundations, and physical well-being and motor development	Domain subscores range from 202 to 298 (and to 293 for physical well-being and motor development) Overall scaled score ranges from 202 to 298	Overall scores are categorized into three performance levels: *Demonstrating readiness*: A child demonstrates foundational skills and behaviors that prepare them for curriculum based on kindergarten standards. *Approaching readiness*: A child demonstrates some foundational skills and behaviors that prepare them for curriculum based on kindergarten standards. *Emerging readiness*: A child demonstrates minimal foundational skills and behaviors that prepare them for curriculum based on kindergarten standards.
PARCC	Grades 3–8, spring[a]	Mastery of grade-specific standards for English language arts/literacy (reading) and math	Overall scaled score for each subject (reading and math) ranges from 650 to 850	Subject scaled scores are categorized into five performance levels: *Level 5*: Exceeded expectations *Level 4*: Met expectations *Level 3*: Approached expectations *Level 2*: Partially met expectations *Level 1*: Did not yet meet expectations

KRA is Kindergarten Readiness Assessment. For more information about this assessment, see Maryland State Department of Education (n.d.) and WestEd (2014). PARCC is Partnership for Assessment of Readiness for College and Careers. For more information about this assessment, see Partnership for Assessment of Readiness for College and Careers (n.d.) and Pearson (2019).

a. End-of-course assessments in English, Algebra I, Algebra II, and Geometry are administered in high school and excluded here because they are not used in MSDE's growth measures.

Note: Reliability estimates for the KRA and PARCC are provided in technical reports. The raw score reliability coefficients for the assessments used in this study are as follows: 0.95 for the 2014/15 KRA overall score, 0.90 for the grade 3 PARCC reading assessment (0.88 for paper-based), and 0.94 for the grade 3 PARCC math assessment (0.93 for paper-based) (Pearson, 2019; WestEd, 2014).

Source: Maryland State Department of Education (n.d.) and Partnership for Assessment of Readiness for College and Careers (n.d.)

2. Sample

In collaboration with MSDE, the study team established rules to determine which schools and students would be used in the analysis. The study used a sample of students who were eligible for inclusion in the K–3 growth measure because they took both the KRA in fall 2014 and the grade 3 PARCC assessment in spring 2018. The study calculated growth measures for the sample of schools that these students attended.

School sample. Schools must have had at least 10 eligible students for growth measures to be reported, per MSDE's Every Student Succeeds Act plan. (Definitions for eligible students are defined below.) Among the 977 schools serving K–3 students, 905 schools had at least 10 eligible students.

Student sample. The analysis for math included 54,393 students; the analysis for reading included 54,397 students. To be included in the analysis, students needed a 2014/15 KRA score and a 2017/18 grade 3 PARCC score in math or reading. Because a 2014/15 KRA score and a 2017/18 PARCC score were required to calculate growth, the analysis necessarily excluded students who were in kindergarten in 2014/15 and then skipped or repeated a grade before grade 3. Similarly, the analysis excluded students who joined the MSDE school system after their cohort's kindergarten assessment window and/or left the system before their cohort's grade 3 assessment window. Students with significant cognitive disabilities who took only the alternate assessment (that is, the Multi-State Alternate Assessment) were also excluded.

Among the original cohort of students who completed the KRA in 2014/15 (that is, the group of students for which growth could have been calculated, if the student also completed the grade 3 PARCC), about 86 percent of the

cohort was included in the analysis. About 14 percent were excluded from the analysis because they lacked a grade 3 assessment score (for example, because they left the Maryland public schools or experienced non-traditional grade progression after taking the KRA). A very small percentage (0.37 percent) were excluded because they took the alternate assessment instead of the PARCC.

C. Calculating the school growth measure

To calculate the school growth measure, the study team first calculated a measure of growth for each student. For each school, we then took the average of the student growth measures for the students who attended that school. The school growth measure therefore provides an estimate of the school's contributions to a typical student's academic growth. We discuss these two steps below.

1. Student growth measure

Student growth measures are intended to measure growth in a way that gives each student an equal opportunity to achieve high growth, regardless of their score on the initial assessment. In other words, it is possible for a student with a low score on the initial assessment to have high growth, and vice versa. Similarly, school growth measures give schools equal opportunity to obtain a high growth estimate, regardless of whether the school serves students with greater or lesser degrees of readiness. This means that all schools have the same chance of obtaining a high growth estimate, regardless of where they started.

Calculating the school growth measure: An overview

1. For each student, calculate a grade 3 test score percentile relative to other students who had the same kindergarten test score (called a student growth percentile, or SGP).
   - Use computer software designed to calculate these percentile scores using advanced statistical methods.
   - Apply statistical methods to adjust for measurement error in kindergarten test scores.
2. Calculate the average grade 3 test score percentile among students in each school (called a mean SGP, or MGP).
   - Account for student movement across schools over time.
3. Calculate confidence intervals, which are needed to identify true differences in growth when comparing schools.

Further details on calculating the school growth measure are in Appendix A.

The study measured student growth using student growth percentiles (SGPs). An SGP is the percentile rank of a student's current score relative to other students who had the same (or a similar) prior score as that student (Betebenner, 2009). For example, a student-level SGP of 75 indicates that the student scored higher than three-quarters of the students who had similar academic achievement at baseline. The study calculated SGP estimates as described in Betebenner (2009).

The study used overall KRA scaled scores to group students based on their academic performance at kindergarten entry, meaning that all students in a particular group had similar KRA scores. The study then assessed each student's performance in grade 3, as measured by PARCC scaled scores, relative to their group at kindergarten entry. SGPs were estimated separately for grade 3 reading and math. Further details on estimating SGPs are in Appendix A.

When calculating SGPs, the study accounted for measurement error in the KRA score. Measurement error means the extent to which KRA scores do not reflect students' actual ability. Measurement error causes some students with high ability to incorrectly receive low KRA scores, and vice versa. The study accounted for measurement error because it can cause SGPs to be too low for students or schools with low prior achievement and too high for students or schools with high prior achievement (Castellano & McCaffrey, 2017; Shang et al., 2015). The study

accounted for measurement error using a method called simulation extrapolation (SIMEX), as described in Shang et al. (2015). Additional details about these analyses are in Appendix A.

2. School growth measure

To calculate each school's growth measure, the study calculated the average SGP—or mean SGP (MGP)—among the students who attended the school. The study calculated mean SGPs, rather than median SGPs, for two reasons: (1) our analyses showed MGPs are more precise than median SGPs (Dragoset et al., 2019) and (2) prior research has shown that MGPs are less biased than median SGPs, meaning they are more likely to reflect schools' actual contribution to students' academic growth (Castellano & Ho, 2015; Castellano & McCaffrey, 2017). The study also calculated a two-year average MGP for each school, by taking the average of the school's MGP from the 2015–2019 cohort and the MGP from the 2014–2018 cohort.

When calculating MGPs, the study accounted for student movement across schools over time. Students who transferred schools between the start of kindergarten and the end of grade 3 contributed to the growth measure for each school they attended, based on the amount of time the student was enrolled in each school. Schools' MGPs were constructed as weighted averages of their students' SGPs, with students who spent more time at the school receiving more weight than students who spent less time at the school. The study accounted for student movement across schools because the opportunity for movement over the K–3 period, which spans four grades, is significantly higher than the amount expected for the grade 4–8 SGPs that MSDE currently calculates, which each span only one grade. More details are in Appendix A.

D. Assessing validity and precision of the school growth measure

The study assessed whether the growth measure is credible—or valid—in the sense that it measures what it is intended to measure: schools' true contributions to their students' K–3 growth. If the measure is not valid, MSDE would not be able to identify elementary schools with low- and high-performing early grades.

The study also assessed the consistency—or precision—of the growth measure, meaning how much it may vary from year to year even if schools' true performance is not changing. If school growth measures are not sufficiently precise, it would be difficult for MSDE to determine whether changes in those measures over time reflect true changes in school performance.

The study assessed the validity and precision of the K–3 school growth measure relative to the growth measures that MSDE uses for accountability in grades 3–6. These included growth measures for grades 3 to 4 and growth measures for grades 3 to 6.

Validity: Is the growth measure credible? Does it appear to be measuring what it is intended to measure: schools' true contributions to their students' K–3 growth? That is, is student academic performance in kindergarten and grade 3 related, as measured by the assessments?

Precision: Is the growth measure consistent? That is, will schools' growth measures vary from year to year even if their true performance is not changing?

1. Validity

In Maryland, students with high KRA scores tended to also have high grade 3 PARCC scores, and students with low KRA scores tended to have low grade 3 PARCC scores (Appendix Exhibits A.6 and A.7). However, to produce valid measures of K–3 growth, performance on the KRA must capture aspects of student academic ability that benefit from K–3 instruction and are measured by performance on the grade 3 PARCC. If the relationship between KRA and grade 3 PARCC scores is weaker than the relationships between PARCC scores for different grade levels, it would suggest that the KRA and grade 3 PARCC are measuring different aspects of academic ability, potentially compromising the measure's ability to accurately measure schools' contributions to student academic growth.

To examine the validity of the K–3 growth measure, the study compared the strength of the relationship between students' KRA scores and grade 3 PARCC scores to the strength of the relationships between students' grade 3 and grade 4 PARCC scores, for three cohorts of students. The study also compared the strength of the relationship between students' KRA scores and grade 3 PARCC scores to the strength of the relationship between students' grade 3 and grade 6 PARCC scores (which involve a similar amount of time between assessments as the K–3 growth measures). The study measured the strength of these relationships by calculating a correlation coefficient, which is a number that can range from -1 to 1, with numbers further from zero indicating a stronger positive or negative relationship.

The study found that schools' K–3 growth estimates are likely less valid than schools' grades 3–4 growth estimates because the correlation between students' KRA and grade 3 scores is significantly lower than the correlation between students' grades 3 and 4 scores (Exhibit 5). The correlation between students' KRA and grade 3 scores is also significantly lower than the correlation between students' grade 3 and 6 scores.

Exhibit 5. Correlation between students' initial and subsequent assessment scores, by cohort

Grades and school years	Correlation coefficient	
	Math	Reading
Between K (2014/15) and grade 3 (2017/18) scores	0.53	0.53
Between grade 3 (2014/15) and grade 4 (2015/16)	0.86	0.82
Between grade 3 (2015/16) and grade 4 (2016/17)	0.87	0.84
Between grade 3 (2016/17) and grade 4 (2017/18)	0.87	0.85
Between grade 3 (2014/15) and grade 6 (2017/18)	0.82	0.77

Note: For each subject (math and reading), the study tested three conditions: (1) whether the correlation between kindergarten (Kindergarten Readiness Assessment [KRA]) and grade 3 (Partnership for Assessment of Readiness for College and Careers [PARCC]) scores (shown in the first row of the table) differed from the correlation between grade 3 and grade 4 PARCC scores (this test was run separately for each of the grade 3–4 cohorts, shown in the second through fourth rows of the table); (2) whether the correlation between grade 3 and grade 4 PARCC scores differed from the correlation between grade 3 and grade 6 PARCC scores shown in the last row of the table (this test was run separately for each of the grade 3–4 cohorts, shown in the second through fourth rows of the table); and (3) whether the correlation between kindergarten and grade 3 scores differed from the correlation between grade 3 and 6 scores. All of these tests were significant at the 5 percent significance level.
Source: Administrative data provided by the Maryland State Department of Education.

The study also examined whether the validity of the K–3 school growth measure would be higher if, rather than using the overall KRA score, it used only the subscores for particular domains. Whereas the PARCC measures students' performance for reading and math, the KRA assesses knowledge and skills in four domains: language and literacy, mathematics, social foundations, and physical well-being and motor development. Some of the domain subscores may relate more closely to grade 3 performance than others, and growth measures will be most valid if they use a configuration of the KRA score that most closely predicts grade 3 PARCC scores. The study found that the overall KRA score performed the best, meaning that it produced a growth measure with higher validity than did the subscores or combinations of subscores.

Although the study found that schools' K–3 growth estimates are likely less valid than schools' grades 3–4 growth estimates, findings from other studies suggest that assessments of young children may be prone to lower correlations with later assessments. A comparison to other studies suggests that the KRA is predicting grade 3 achievement reasonably well, relative to other kindergarten assessments. For example, a range of kindergarten readiness measures had weak to moderate correlations with grade 3 achievement scores, ranging from 0.10 to 0.61 for grade 3 reading and 0.12 to 0.68 for grade 3 math, in two large United States studies (Duncan et al., 2007). The highest correlations from these studies are from a measure of math skills at kindergarten entry that was designed to track development over time in the 1998 Early Childhood Longitudinal Study, and these correlations are only slightly higher than the correlations found between KRA and grade 3 PARCC scores.

2. Precision

All school growth measures are *estimates* of schools' true performance and may not equal true performance for various reasons. For example, a high-performing student's test score might not reflect the student's true abilities if they are feeling ill when they take the test. Therefore, a school's growth measure based on that student's test score might not equal the schools' true contribution to student academic growth.

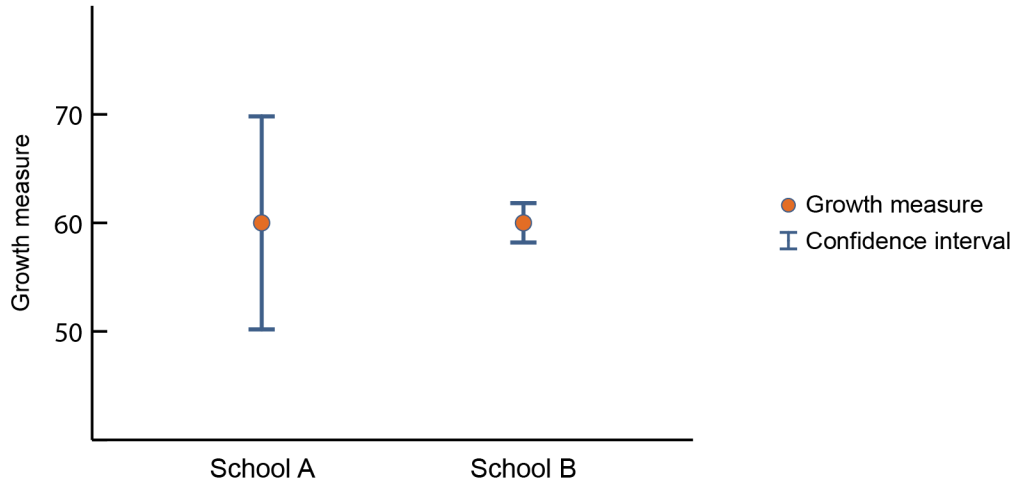
The precision of a school's growth estimate (such as an MGP) indicates how confident we are that the growth estimate equals the school's true contribution to student academic growth. If school growth measures are not sufficiently precise, it will be difficult to determine whether changes in those measures over time reflect true changes in school performance.

What is a 95 percent confidence interval?

This interval is a range of values that would contain the school's true MGP 95 percent of the time, if the MGP calculation was repeated many times using different random samples of students in the school.

To examine precision, the study team calculated a confidence interval around each school's growth measure. A confidence interval is a set of two numbers—one less than the school's growth measure and one greater than it. These two numbers are called the upper and lower confidence limits. For example, a particular school—call it School A—might have a growth measure (that is, MGP) of 60 and a confidence interval of 50 to 70. The width of the confidence interval indicates to what extent the growth estimate equals the true change in school performance. Narrower confidence intervals are more desirable and indicate more precision. Wider confidence intervals indicate less precision. Continuing with the example described above, School A's growth measure has a confidence interval width of 20 (range = 50 to 70). Suppose another school—call it School B—has a growth estimate of 60 and a confidence interval of 58 to 62, or a confidence interval width of 4 (range = 58 to 62). School B's growth measure is more precise than School A's growth measure because its confidence interval is narrower (Exhibit 6). More details about how the study calculated a 95 percent confidence interval for each school's growth measure are in Appendix A.

Exhibit 6. School B's growth measure is more precise than School A's growth measure because its confidence interval is narrower

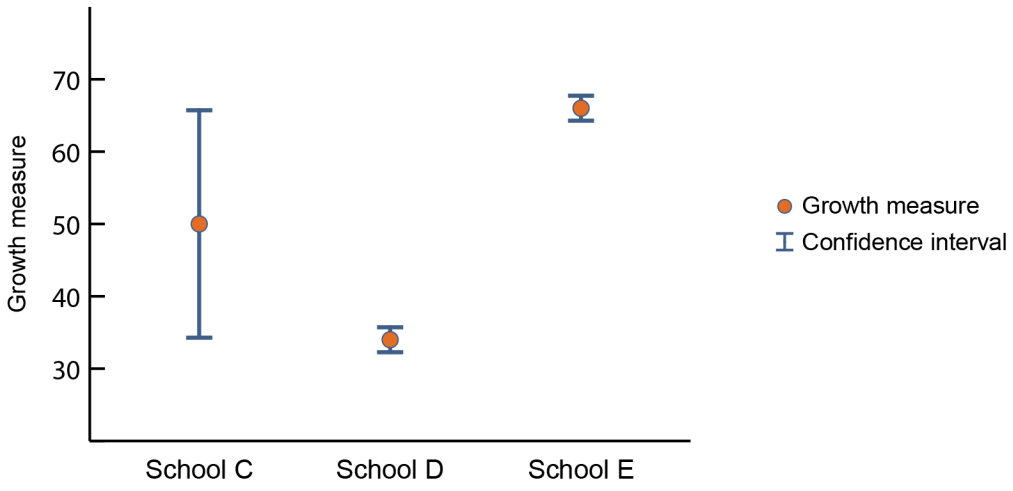


Source: Hypothetical data.

Two schools with different growth estimates cannot be considered to have true differences in growth unless their confidence intervals do not overlap (Exhibit 7). For example, suppose an average school—that is, a school with an MGP of 50—has a confidence interval of 32; call this School C. We could not conclude that School C has a different

level of growth than another school—call it School D—that has an MGP of 34 (16 percentile points below 50) or a third school—call it School E—that has an MGP of 66 (16 percentile points above 50).

Exhibit 7. Schools D and E cannot be considered to have true differences in growth from School C because their confidence intervals overlap with School C's confidence interval



Source: Hypothetical data.

The study assessed the precision of the K–3 school growth measure relative to an existing school growth measure in Maryland's accountability system. In particular, the study compared the average confidence interval width for schools' K–3 growth measures to the average confidence interval width for schools' grades 3–4 growth estimates (calculated using the SGP and MGP methods described above).

The study found that schools' K–3 growth measures had a similar level of precision as schools' grades 3–4 growth measures. The average confidence interval width for schools' K–3 growth measures was 12 percentile points for math and 13 for reading, compared to 13 for schools' grades 3–4 growth measures in both math and reading (Dragoset et al., 2019).

The study also found that schools' K–3 growth measures are much less precise for smaller schools than for larger schools. For math, the average confidence interval width was 32 percentile points for the smallest schools (those with fewer than 15 tested students) and 8 percentile points for the largest schools (those with more than 180 tested students; Appendix Exhibit A.8). Results for reading were similar.

Lastly, the study found that administering the KRA to a subset of students in each classroom greatly reduces the precision of schools' K–3 growth measures. Beginning in school year 2016/17, Maryland law allowed districts to administer the KRA to a random subset of students in each classroom. Statewide, 34 percent of students took the KRA in 2016/17, 35 percent took it in 2017/18, and 39 percent took it in 2018/19. As a result, the average width of confidence intervals around schools' K–3 growth measures doubled from roughly 12 to roughly 25 percentile points (Appendix Exhibits A.9 and A.10).

E. Strengths and limitations of the school growth measure

The key strength of the K–3 school growth measure calculated in the study is that it enables MSDE to identify schools with low- and high-performing early grades, and inform policy and resources to improve those schools. Prior to the calculation of the measure, there was a lack of standardized, consistent information on schoolwide student growth for the early elementary grades. The growth measure can help MSDE and local school officials identify areas of strength or need that warrant further investigation, by knowing how well each school is

supporting the academic growth of its youngest students. We discuss the limitations of the school growth measure below.

States need to consider the validity of K–3 growth estimates before they are used for school accountability. This study identified some concerns about validity. In particular, the study's findings suggest that schools' K–3 growth estimates are likely less valid than schools' grades 3–4 growth estimates. If K–3 estimates are prone to be less valid than later grade estimates, states could acknowledge these differences by assigning less weight to the K–3 growth measure in their accountability framework, relative to growth measures in later grades. States could also consider publicly reporting the K–3 growth measure, but not using it in their formal accountability system.

The K–3 school growth measures calculated in the study reflected only the students who had a KRA score and a grade 3 PARCC score, both of which were required to calculate growth. Students included in the analysis had higher KRA scores and higher PARCC scores than the students who were excluded from the analysis (Dragoset et al., 2019). If growth could be calculated for students missing one or both scores, growth measures might change for the schools in which these students were enrolled.

The findings and conclusions may also differ under different growth models, such as a value-added models, which other states use to calculate growth. SGPs do not adjust for differences in student characteristics other than prior achievement.

The study could not measure year-to-year stability (that is, reliability) for schools' K–3 growth estimates because the 2017/18 grade 3 cohort was the first cohort to have completed the KRA. Examining year-to-year reliability can boost schools' confidence in the measure by demonstrating that estimates are not driven by random year-to-year fluctuation in the data (Goldschmidt, Choi, & Beaudoin, 2012).

Finally, the growth estimates calculated in the study could not account for student movement from one school to another *within* school years because such movement could not be observed in the data. Because the data only listed a single school for each student in each school year, the growth estimates account for student movement *between* school years, but not *within* school years. Collecting and maintaining data on all the schools that students attend during each school year would enable states to account for this type of movement when estimating schools' K–3 growth.

F. Looking to the future

To effectively guide policy and resources, there is an emerging need for states to have information about how well their elementary schools are supporting student learning in all grades. This memo describes an approach to measuring schools' contributions to student academic growth in the early grades when limited statewide assessments are available. MSDE and REL Mid-Atlantic collaborated to calculate K–3 growth using two different assessments and provided a model to other states interested in constructing similar measures. Maryland plans to continue administering the KRA in future school years and strengthen the elementary school growth measure with additional state assessments in the early grades. Additional research could examine K–3 growth estimates based on multiple cohorts, to help improve the validity and precision of the estimates.

APPENDIX A. Methods and additional findings

This appendix provides details about the methods used to calculate the K–3 school growth measure and assess its validity and precision and presents additional findings.

A. Calculating student growth percentiles

Estimating a student growth percentile (SGP) involves two steps: (1) estimating quantile regressions of students' current scores on their prior scores, and (2) calculating students' SGP estimates using the results from those regressions. In the first step, the study estimated quantile regressions (as described in Koenker, 2005) of students' current scores on their prior scores.[1] Let Q_i^τ denote the τth quantile of the current score, given student i's value of the prior score, where i runs from 1 to the number of students (n). For example, for τ = 0.50 (the median), Q_i^τ denotes the median current score among all students with a particular prior score. In the second step, the study compared each student's current score to the τ distinct values of Q_i^τ, found the two consecutive values of Q_i^τ that surround the student's current score, took the midpoint of those two τ values, and multiplied that by 100. For example, if student i's current score was between $Q_i^{0.935}$ and $Q_i^{0.945}$, that student's estimated SGP was 94.

The study estimated SGPs using the SGP package (Betebenner et al., 2019) in the statistical software program R (R Development Core Team, 2019). In that package, Q_i^τ is, by default, estimated for τ = 0.005, 0.015, . . ., 0.995. The package estimates Q_i^τ with nonparametric quantile regression, specifically, quantile regression with cubic B-splines. Unlike parametric models that estimate a single regression for the entire dataset, a regression B-spline divides the dataset into multiple bins of students with similar prior scores and estimates a separate regression within each bin (De Boor, 2001; Hardle et al., 2004). In the SGP R package, the type of regression that is estimated within each bin is a cubic polynomial (meaning that current scores are regressed on prior scores, prior scores squared, and prior scores cubed). The points that divide the data into bins are called knots. The regression functions on either side of each knot are constrained to have the same value and slope at the knot where they meet, so that the overall regression function is smooth at the knots. In the SGP R package, by default, the knots are defined as the 20th, 40th, 60th, and 80th percentiles of the prior score.

B. Accounting for measurement error in the Kindergarten Readiness Assessment score

Measurement error in students' prior scores results in biased SGP estimates for students and biased mean SGP (MGP) estimates for schools (Castellano & McCaffrey, 2017; Shang et al., 2015). In particular, SGP and MGP estimates tend to be underestimated for students or schools with low prior achievement and overestimated for students or schools with high prior achievement. Based on consultations with the Maryland State Department of Education (MSDE), the K–3 model adjusted for measurement error in students' prior scores using the SIMEX method.

The SIMEX method is a measurement error correction technique, originally proposed by Cook and Stefanski (1994), that reduces this bias in SGP and MGP estimates (Castellano & McCaffrey, 2017; Shang et al., 2015). The technique first uses simulation to measure the consequences for the quantile score estimates, Q_i^τ, of adding incrementally more measurement error to the prior assessment scores. Then, the technique uses the relationship between the amount of measurement error and the quantile score estimates to extrapolate how the quantile

[1] Whereas a linear regression of students' current scores on their prior scores estimates the average change in the current score associated with a one unit increase in the prior score, a quantile regression of students' current scores on their prior scores estimates the change in a specified quantile of the current score associated with a one-unit increase in the prior score. For example, a median regression (the median is the 50th percentile) of students' current scores on their prior scores estimates the change in the median current score associated with a one-unit increase in the prior score.

estimates would appear had there been no measurement error in the prior assessment scores. Executing the method consists of two steps: simulation and extrapolation.

In the simulation step, an increasing amount of extra measurement error is added to students' Kindergarten Readiness Assessment (KRA) scores, and the quantile scores are calculated for each level of error. This study used four levels of error, described in more detail below. For each of those four levels of measurement error, the following process was repeated. First, a distribution, or set of likely values, for the measurement error was chosen to correspond to the level of measurement error, so that larger errors were more likely when the amount of error was higher. Second, a particular random amount of error from that distribution was chosen for each student and added to that student's KRA score repeatedly 100 times, producing 100 new datasets of students' prior assessment scores. Third, using each of the 100 datasets, quantile scores Q_i^τ were then calculated using the standard method described above. Lastly, the quantile scores were averaged across the 100 iterations, resulting in a single set of quantile scores for the level of measurement error. The study assumed measurement error was distributed according to a normal distribution with a mean of zero and used a value of 10.14 for σ^2, the variance of the measurement error in the observed prior assessment score data. This value was the variance of measurement error for the 2014 overall KRA score based on the reported standard error of measurement and reliability coefficient (WestEd, 2014). The variance of the measurement error added to students' scores for each level of simulated measurement error was equal to $\lambda\sigma^2$, where λ was a number chosen between 0 and 2. For each level of simulated measurement error, the total measurement error (original measurement error plus added measurement error) had a variance equal to $(1 + \lambda)\sigma^2$. This study used four values of λ, set to 0.5, 1, 1.5, and 2, following Carroll et al. (1999).

In the extrapolation step, the relationship between the averaged quantile scores for each level of measurement error and the value of λ corresponding to that level was measured and used to understand how the quantiles might appear, had no measurement error been present. The relationship was measured using an ordinary least squares regression, and the coefficient from that regression was used to calculate the predicted value of each quantile score at λ = -1, which is the level of λ that corresponds to a measurement error variance of zero (because $(1 + \lambda)\sigma^2$ is equal to 0 when λ = -1). These predicted values are the SIMEX estimates of the quantile scores, denoted as $Q_{i,SIMEX}^\tau$. The SGP estimate derived from $Q_{i,SIMEX}^\tau$ was denoted as SGPSIMEX.

Following McCaffrey et al. (2015), the study made one final adjustment to the SIMEX-adjusted SGP estimates, to improve the distributional properties of those estimates. In particular, the study took the percentile ranks of the SIMEX-adjusted SGP estimates, which was an ordinal transformation that preserved the ordered ranking of students' SGP estimates. McCaffrey et al. demonstrated that SIMEX-adjusted SGPs have a U-shaped distribution, meaning that larger percentages of students have SGP values near 0 and 100 than near 50. In contrast, non-SIMEX-adjusted SGPs, as well as SGPs based on an assessment without any measurement error, have the appealing property of being uniformly distributed (meaning that the percentage of students with a particular SGP value is the same for every value). McCaffrey et al. showed that taking the percentile ranks of SIMEX-adjusted SGPs results in SGP estimates that are uniformly distributed and that retain the desirable property of SIMEX-adjusted SGPs (that is, they reduce the bias caused by measurement error in the prior assessment).

As noted on page 6, school-level growth estimates were then constructed as weighted averages of their assigned students' SGPs.

C. Accounting for student movement between schools

To account for student movement between schools, the study assigned weights proportional to the amount of time the student was enrolled at each school over the K–3 period. Proportional weighting comprised three steps: (1) identifying each school that the student attended over the period, (2) determining the proportion of

the K–3 period that the student was enrolled at each school, and (3) constructing the final weights as described below:

1. Identifying schools. Ideally, every school that the student attended between administration of the KRA and the grade 3 Partnership for Assessment of Readiness for College and Careers (PARCC) would receive credit for a portion of the student's academic growth over this period (that is, the student's SGP estimate). However, a complete attendance record for each student was not part of the data available for this study. Rather, students' school enrollment in the dataset was observed at six points in time: (1) at KRA administration, (2) on the last day of kindergarten, (3) on the last day of grade 1, (4) on the last day of grade 2, (5) at PARCC administration, and (6) on the last day of grade 3.

Because the data provided by MSDE only listed a single school for each student in each school year, the growth measures accounted for student movement between school years, but not within school years. Collecting and maintaining data on student movement within a school year, such that there is attendance data on all schools that students attend during each school year, would enable states to account for this type of movement when estimating schools' contributions to students' academic growth.

The student attendance data identified the school where the student was enrolled as of the last day of the school year in kindergarten through grade 3 (hereafter referred to as the student's end-of-year school), the date the student enrolled in the school that year, and the number of days the student was in attendance and absent. The assessment data identified the school where the student was enrolled when the student completed the KRA and grade 3 PARCC (hereafter referred to as the student's assessment school). Multiple transfers between these two time points and attendance at other schools before the assessments each year were not observed.

To calculate schools' growth measures, students were assigned to the single school in each school year for which the number of days they were in attendance or absent was available: their end-of-year school. The one exception to this rule is that students were assigned to their grade 3 assessment school for grade 3, even when that school differed from their end-of-year school that year. In the latter case, students transferred to their end-of-year school after taking the grade 3 assessment, so the end-of-year school did not contribute to the student's academic growth as measured by the grade 3 assessment. Such students were therefore not assigned to their end-of-year school for the grade 3 school year to prevent them from contributing to a growth measure for a school they did not attend between the KRA and grade 3 PARCC. This included instances where the student had a grade 3 assessment score but was not observed in the grade 3 attendance data file, indicating the student exited the MSDE school system after taking the PARCC.

2. Calculating the percentage of time enrolled. Students' percentage of days enrolled in their assigned school for each year was defined as the number of days they were enrolled in the school that year out of the total number of days the school was in session the same year. The number of days enrolled was defined as the sum of the total days attended and the total days absent, as indicated for student's end-of-year school on the attendance data provided by MSDE. The total number of days the school was in session was defined as 180 days, which is the number of instructional days for the MSDE school year.

Because the student attendance data only reported days attended and days absent for the student's end-of-year school, data on students' total days enrolled in the assessment school were not available when the school differed from their end-of-year school. In these cases (less than 1 percent of students in the K–3 sample), the percentage of days enrolled in the assessment school was therefore estimated and assumed the student was enrolled in the assessment school for the portion of the year that the student was not enrolled in the end-of-year school. That is, when the two schools differed, the percentage of days enrolled at the grade 3 assessment school was calculated as 100 percent minus the percentage of days enrolled in the grade 3 end-of-year school.

3. Constructing proportional weights. The SGP for each of the school's students was weighted according to the amount of time the student was enrolled in the school over the four-year period. Specifically, each SGP was apportioned to the student's assigned school in each year with a weight equal to the proportion of time the student was enrolled in the school that year multiplied by 0.25, to reflect that growth was estimated over four years. Students' growth measures could be more precisely apportioned across schools by using more comprehensive student enrollment data, including whether a student attended part-day or full-day kindergarten, and with detailed information on the timing of the PARCC in each school.

For example, a student who enrolled in kindergarten at the start of the year and attended the same school—named School A—through grade 3, without ever transferring schools, would be observed with 100 percent days enrolled at the same school in each of the four years, as indicated in Exhibit A.1. Therefore, this student's SGP would be entirely attributed to School A.

Exhibit A.1. Example of constructing proportional weights when a student does not move schools between kindergarten and grade 3

Grade	School	Percent of days enrolled	Percent of four-year period	Weight
K	A	100	100/4=25	1 * 0.25 = 0.25
1	A	100	100/4=25	1 * 0.25 = 0.25
2	A	100	100/4=25	1 * 0.25 = 0.25
3	A	100	100/4=25	1 * 0.25 = 0.25

Source: Hypothetical data.

Next, take an example of a student who enrolled in kindergarten at the start of year and attended the same school (School B) through grade 2, but then changed schools (to School C) *between* grades 2 and 3. This student would be also observed with 100 percent days enrolled in each of the four years, but their school would change in grade 3. Therefore, 75 percent of the student's SGP would be attributed to School B, which taught the student for three full years, and 25 percent would be attributed to School C, which taught the student one full year (Exhibit A.2).

Exhibit A.2. Example of constructing proportional weights when a student changes schools between grades 2 and 3

Grade	School	Percent of days enrolled	Percent of four-year period	Weight
K	B	100	100/4=25	1 * 0.25 = 0.25
1	B	100	100/4=25	1 * 0.25 = 0.25
2	B	100	100/4=25	1 * 0.25 = 0.25
3	C	100	100/4=25	1 * 0.25 = 0.25

Source: Hypothetical data.

Next, take an example of a student who enrolled in kindergarten at the start of the year and then changed schools (from School D to School E) in the middle of grade 1. This student would be observed with 100 percent days enrolled in kindergarten for School D, 100 percent days enrolled in grades 2 and 3 for School E, but only 50 percent days enrolled in grade 1, because the student enrolled in their end-of-year school (School E) in the middle of the school year (Exhibit A.3). School E, which taught the student for two and a half years, would receive 62.5 percent (or 2.5 years divided by 4 years) of the student's SGP. School D, which taught the student for one and a half years, should receive the remaining 37.5 percent. However, it would receive only 25 percent, and the remaining 12.5

percent would not be attributed to any school's growth measure, because the student's enrollment in School D during the first half of grade 1 is unobserved in the data.

Exhibit A.3. Example of constructing proportional weights when a student changes schools in the middle of grade 1

Grade	School	Percent of days enrolled	Percent of four-year period	Weight
K	D	100	100/4=25	1 * 0.25 = 0.25
1	E	50	100/4=25	0.5 * 0.25 = 0.125
2	E	100	100/4=25	1 * 0.25 = 0.25
3	E	100	100/4=25	1 * 0.25 = 0.25

Source: Hypothetical data.

Finally, take an example of a student who enrolled in kindergarten at the start of the year, left the MSDE school system before the end of the year, but then returned in grade 1 and attended the same school (School F) through grade 3. This student would be observed with zero days enrolled in kindergarten (because the student did not have an end-of-year school), and then 100 percent days enrolled in grades 1 through 3 (Exhibit A.4). School F would receive 75 percent of the student's SGP, and the remaining 25 percent would not be attributed to any school because the school the student attended during kindergarten was unobserved in the data.

Exhibit A.4. Example of constructing proportional weights when a student leaves the school system in the middle of kindergarten and returns in grade 1

Grade	School	Percent of days enrolled	Percent of four-year period	Weight
K	N/A	0	0/4=25	0 * 0.25 = 0
1	F	100	100/4=25	1 * 0.25 = 0.25
2	F	100	100/4=25	1 * 0.25 = 0.25
3	F	100	100/4=25	1 * 0.25 = 0.25

N/A = not available.
Source: Hypothetical data.

These examples illustrate two important points. First, in this study the credit a school received for a student's growth (that is, the student's SGP) was proportional to the amount of time the student was (observed to be) enrolled in the school over the four-year period, with each of the four years weighted equally. Second, schools only received credit for observed enrollment, and a student's enrollments could not be completely observed in the data if the student transferred schools during a school year and/or left and then returned to the MSDE school system between the K–3 period.

D. Calculating confidence intervals for schools' growth estimates

The study calculated a 95 percent confidence interval for each school's growth measure using a method known as bootstrapping. This method involves recalculating schools' growth estimates many times using different random samples of students from the original data and then examining the distribution of growth estimates calculated from those many samples. The bootstrapping method was necessary because the method used to calculate the growth measures (described in an earlier section) does not allow a more straightforward precision calculation. Below is an overview of the bootstrapping method, followed by details of how it was implemented for the K–3 growth estimates and the grades 3–4 growth estimates.

Overview of the bootstrapping method. Bootstrapping is a method for calculating a measure of variance or precision (such as a standard error or a confidence interval) for a sample statistic (such as a mean or a median). The standard error of any sample statistic is the standard deviation of the distribution of a large number of such statistics derived from the larger population of interest (called the sampling distribution for the statistic). For example, if one calculated the median height of a sample of women from the general population, the standard error of that median is a measure of how much that median value would differ across many different samples.[2] One can approximate the population-wide sampling distribution for a statistic by repeatedly resampling observations from the sample in the original data. In other words, instead of drawing many different samples from the population, one can draw many different samples from the original sample, with replacement (meaning that each individual observation can be selected multiple times), and then calculate the standard deviation of the statistic across those many samples. This process of resampling from the original data many times and then examining the distribution of a statistic calculated from those many samples is called bootstrapping. The number of samples is referred to as the number of bootstrap replications.

Implementing the bootstrapping method for K–3 growth estimates. To calculate confidence intervals for schools' K–3 growth estimates, each bootstrap replication consisted of the following three steps:

1) Restrict the sample to students eligible for the particular analysis (for example, the math analysis or the reading analysis) and then randomly resample students within schools, with replacement. Sampling was done within schools to ensure that the size of each school (measured as the number of students contributing information to the school's growth estimate) was constant across bootstrap replications, so that each school's confidence interval would reflect variation in *which* students attended the school, but not *how many* students attended the school. More specifically, the study team sampled students within schools using a file containing all unique student–school combinations that appeared in the original data and their associated weights. As a simplified example, imagine an original dataset containing three students, in which the first student attended School A for all four years from 2014/15 through 2017/18, the second student attended School B for all four years, and the third student attended School A for two years and School B for two years. In this example, the first step of the bootstrap replication would randomly resample students within Schools A and B from a file that looks like Exhibit A.5.

Exhibit A.5. Example of file containing students within schools

Student	School	Weight
1	A	1
2	B	1
3	A	0.5
3	B	0.5

Note: Student 3 is listed twice—for both School A and School B—because this student attended both of these schools during the four-year period of interest; they attended School A for the first two years and then attended School B for the last two years.
Source: Hypothetical data.

2) Calculate SGP estimates as described above in the "Calculating SGPs" section. This step included two noteworthy items:
   a) For the purpose of calculating these SGP estimates—which did not feed into the school growth estimates that were the focus of this study, but which were simply temporary SGP estimates that were used to calculate *confidence intervals* around schools' growth estimates—the dataset of resampled students that

[2] Imagine repeating the exercise 1,000 times, each time drawing a different sample of women and then saving the median value calculated from that sample. The standard error of the median is equal to the standard deviation of the medians calculated for those 1,000 samples.

came out of step 1 was restricted to the set of unique students among those sampled. In other words, students picked twice only contributed one observation to the SGP calculation, rather than two. Students who switched schools during the study period were randomly selected more often than students who did not because they appeared more times in the file from which students were randomly resampled (in the example file above, student #3 [who switched schools during the study period] appeared more often than the other students [who did not switch schools]). Thus, restricting to the set of unique students among those sampled helped ensure that more mobile students (who were likely lower-growth students) were not overweighted in the calculation of SGP estimates simply because they were more mobile than other students. Each student's SGP estimate from this step was then attached to all of that student's observations in the full resampled dataset, and that dataset (in which some students appeared more than once) was used in step 3 below to calculate schools' growth estimates.

b) Although the growth estimates produced by this study were SIMEX-adjusted to account for measurement error in students' prior scores, the study did not make this adjustment when calculating confidence intervals around those growth estimates in order to reduce computational complexity.[3] The confidence intervals produced in this study might be slightly different from those produced using the SIMEX method. On the other hand, the width of the confidence interval around a school's growth estimate is largely driven by school size (that is, the number of students who contribute information to the school's growth estimate), and that sample size would not change if the SIMEX method were used. Thus, the confidence intervals produced by this study are likely a reasonable reflection of the amount of imprecision associated with schools' growth estimates.

3) Calculate each school's growth estimate as the weighted mean of SGP estimates across all observations for that school in the full resampled dataset (in which some students appeared more than once).

At the end of the 100 replications, for each school, the study calculated 2.5th and 97.5th percentiles of the 100 values of the school's growth estimate.[4] These percentiles were the bootstrapped estimates of the upper and lower bounds of the 95 percent confidence interval of the growth estimate for that school. The width of the confidence interval for the school's growth estimate equaled the difference between the upper and lower bounds.

Implementing the bootstrapping method for grades 3–4 growth estimates. To calculate confidence intervals for schools' grades 3–4 growth estimates, the study used the same procedure described above, with three changes. First, the sampling of students within schools was done using a file containing a single record for each student, in which each student was associated with their grade 4 school. Second, for the purpose of calculating SGP estimates in step 2 of each bootstrap replication, rather than using only one record for each unique student as was done for the K–3 growth estimates, students who were sampled repeatedly contributed their observation to the SGP calculation once for each time they were sampled. Unlike the approach used for the K–3 growth estimates, it was not necessary to restrict to unique students because each student appeared in the dataset used for sampling only once (so that the probability of a student being sampled repeatedly did not depend on whether they transferred between schools). Third, no weights were used in calculating schools' growth estimates. Rather, each student's

[3] In this study, a single run of the SGP package on a Windows server took 4 to 10 minutes (depending on various inputs, such as the sample size) with the SIMEX option turned off, and 7 to 15 hours with the SIMEX option turned on. Thus, taking into account the eight models the study team conducted (the model for K–3 growth, three models for grade 3–4 growth using three different cohorts, and two subjects [math and reading] for each of those four models), the team estimated that using the SIMEX option when bootstrapping would take more than 90 days, even when using multiple computer processors simultaneously.

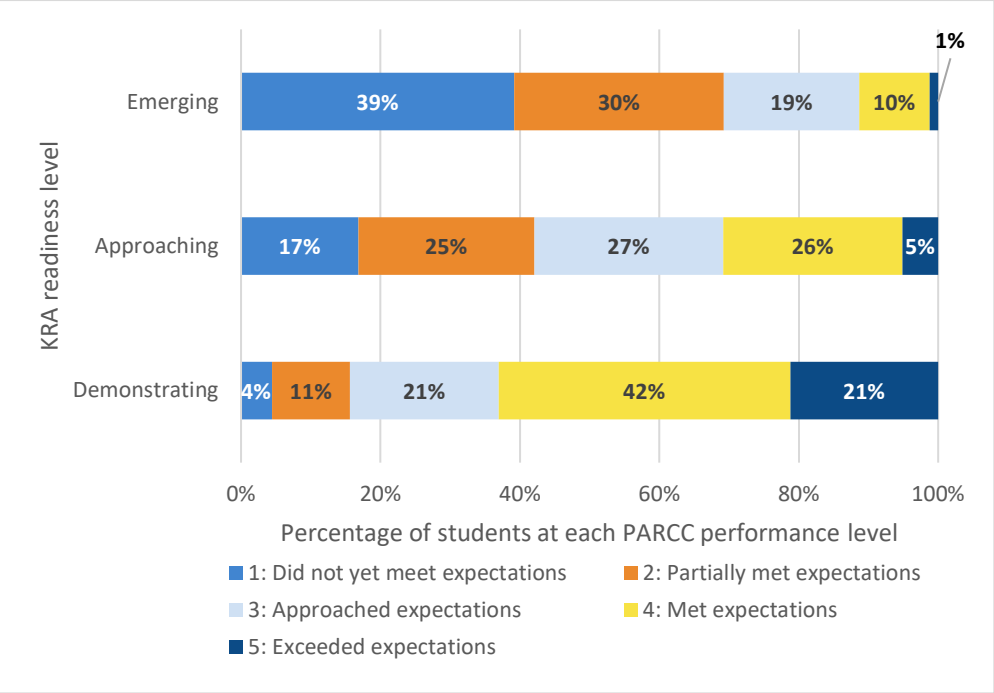
[4] To ensure that confidence intervals could be calculated within a reasonable amount of time, this study used 100 bootstrap replications. To understand the extent to which results might change if the analysis used 1,000 replications, the study ran one of the models (the math K–3 model) using both 100 and 1,000 replications. Results were similar, differing only in the decimal places. Specifically, the average confidence interval width (in percentile points) was 12.33 for schools using 100 replications, compared to 12.49 using 1,000 replications.

full weight (of 1) was given to the school with which they were associated for grade 4, because the grade 4 assessment was designed to measure growth that occurred during grade 4, and because the majority of a student's time between the two assessments was spent in grade 4.

E. Additional findings

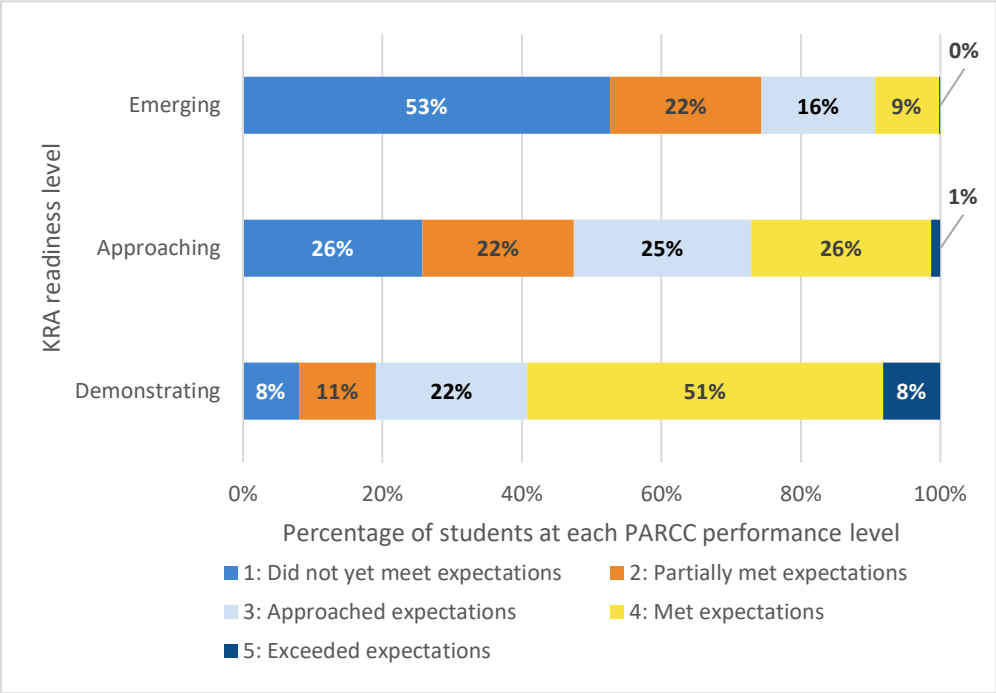
The percentages of students at each PARCC performance level (for math and reading), by KRA readiness level, are shown in Exhibits A.6 and A.7. For example, 12 percent of students who had "emerging" math skills on the KRA met or exceeded expectations on the grade 3 PARCC, versus 63 percent of students who were already "demonstrating" math skills.

Exhibit A.6. Percentage of students at each grade 3 Partnership for Assessment of Readiness for College and Careers math performance level, by Kindergarten Readiness Assessment readiness level



KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

Exhibit A.7. Percentage of students at each grade 3 Partnership for Assessment of Readiness for College and Careers reading performance level, by Kindergarten Readiness Assessment readiness level



KRA is Kindergarten Readiness Assessment. PARCC is Partnership for Assessment of Readiness for College and Careers.
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

The study found that schools' K–3 growth estimates are much less precise for smaller schools than for larger schools (Exhibit A.8).

Exhibit A.8. Average confidence interval width for schools' K–3 growth estimates, by school size

School size (percentile/range)[a]	Average confidence interval width for schools of that size[b]
Math	
1st/1–14 students	32
5th/15–30 students	19
10th/31–41 students	16
25th/42–59 students	14
50th/60–79 students	12
75th/80–107 students	11
90th/108–132 students	10
95th/133–148 students	9
99th/149–179 students	9
100th/180–208 students	8
Reading	
1st/1–14 students	27
5th/15–30 students	20
10th/31–41 students	17
25th/42–58 students	15
50th/59–80 students	13
75th/81–107 students	11
90th/108–132 students	10
95th/133–148 students	10
99th/149–179 students	9
100th/180–209 students	8

a. This column shows school size measured as the number of students contributing information to the school's K–3 growth estimate.
b. This column shows the average confidence interval width for schools in a particular quantile of school size. For example, the first row of the table shows the average confidence interval width for schools in the 1st percentile of school size (that is, schools that have 14 or fewer students contributing information to their growth estimate).
Source: Administrative data provided by the Maryland State Department of Education, 2014/15 to 2017/18.

The study also found that administering the KRA to a subset of students in each classroom greatly reduces the precision of schools' K–3 growth measures (Exhibit A.9).

Exhibit A.9. Average confidence interval width for schools' K–3 growth estimates, by percentage of students who take the Kindergarten Readiness Assessment (KRA)

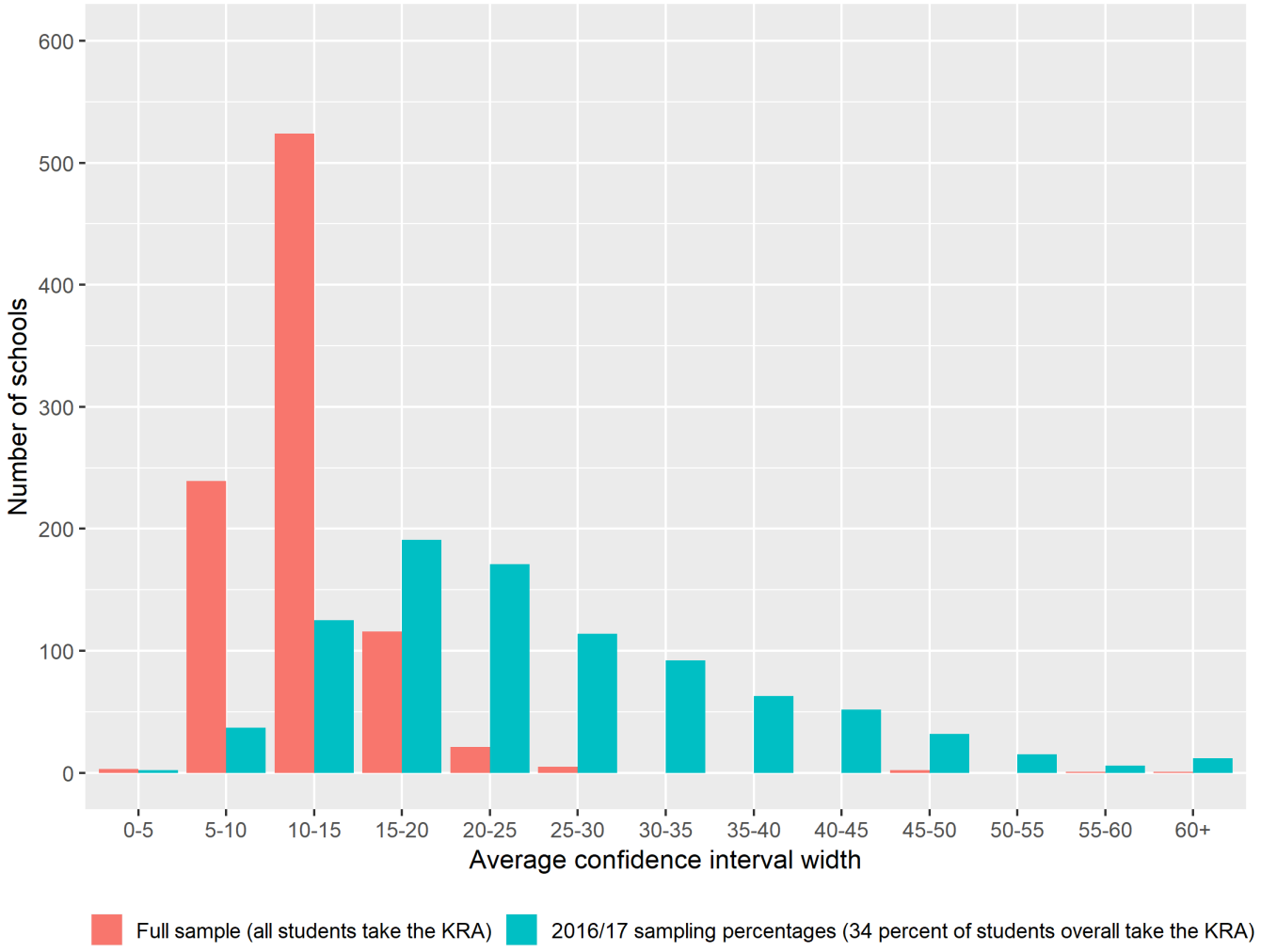
	Average confidence interval width (percentile points)	
Sampling percentage	Math	Reading
2014/15 full sample (all students take the KRA)	12	13
2016/17 sampling percentages (34 percent of students overall take the KRA)	25	26
2017/18 sampling percentages (35 percent of students overall take the KRA)	24	24
2018/19 sampling percentages (39 percent of students overall take the KRA)	23	23

Source: Administrative data provided by the Maryland State Department of Education.

When all students take the KRA, the vast majority of schools have a confidence interval width less than 20 percentile points (Exhibit A.10 shows results for math using the 2016/17 sampling percentages; results for reading and for other sampling percentages were similar). In contrast, when only a third of students take the KRA, more than half of schools have a confidence interval width greater than 20.

Exhibit A.10. Distribution of average confidence interval widths for schools' K–3 math growth estimates, by whether all students or a subset of students take the Kindergarten Readiness Assessment (KRA)

Number of schools



Note: This figure shows results using the same sampling percentages that the Maryland State Department of Education (MSDE) used in the 2016/17 school year.

Source: Administrative data provided by MSDE.

References

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://eric.ed.gov/?q=Norm-+and+criterion-referenced+student+growth&id=EJ866087

Betebenner, D., VanIwaarden, A., Domingue B., & Shang, Y. (2019). *SGP: An R package for the calculation and visualization of student growth percentiles and percentile growth trajectories* (R package version 1.9–0.0). http://centerforassessment.github.io/SGP/

Carroll, R. J., Maca, J. D., & Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika, 86,* 541–554.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics, 40*(1), 35–68. https://eric.ed.gov/?q=Practical+differences+among+aggregate-level+conditional+status+metrics&id=EJ1049833

Castellano, K. E., & McCaffrey, D. F. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues and Practice, 36*(1), 14–27. https://eric.ed.gov/?q=The+accuracy+of+aggregate+student+growth+percentiles+as+indicators+of+educator+performance&id=EJ1135072

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association, 89,* 1314–1328.

De Boor, C. (2001). *A practical guide to splines*. Springer.

Dragoset, L., Baxter, C., Dotter, D., & Walsh, E. (2019). *Measuring school performance for early elementary grades in Maryland*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. https://eric.ed.gov/?id=ED601956

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://eric.ed.gov/?q=School+readiness+and+later+achievement&id=EJ779938

Education Commission of the States. (2020). *50-state comparison: State K–3 policies*. https://reports.ecs.org/comparisons/state-k-3-policies-05

Goldschmidt, P., Choi, K., & Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Washington, DC: Council of Chief State School Officers. Retrieved from ERIC database, https://eric.ed.gov/?q=Growth+model+comparison+study%3a+Practical+implications+of+alternative+models+for+evaluating+school+performance&id=ED542761.

Hardle, W., Muller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models.* Springer.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Maryland State Department of Education. (2018). *Maryland Every Student Succeeds Act (ESSA) consolidated state plan final.* http://marylandpublicschools.org/about/Pages/DAPI/ESSA/index.aspx

Maryland State Department of Education. (n.d.). *Student report: Kindergarten Readiness Assessment.* http://www.marylandpublicschools.org/about/Documents/DAAIT/KRA/KRAISR.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://eric.ed.gov/?q=The+impact+of+measurement+error+on+the+accuracy+of+individual+and+aggregate+SGP&id=EJ1054132

Partnership for Assessment of Readiness for College and Careers. (n.d.) *English language arts/literacy and mathematics sample score reports.* https://osse.dc.gov/sites/default/files/dc/sites/osse/page_content/attachments/PARCC%202019%20Sample%20ELA%20Student%20Score%20Report.pdf and https://osse.dc.gov/sites/default/files/dc/sites/osse/page_content/attachments/PARCC%202019%20Sample%20Math%20Student%20Score%20Report.pdf

Pearson. (2019). *Final technical report for 2018 administration*. https://files.eric.ed.gov/fulltext/ED599198.pdf

R Development Core Team. (2019). *R: A language and environment for statistical computing* (3-900051-07-0). R Foundation for Statistical Computing.

Shang, Y., VanIwaarden, 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://eric.ed.gov/?q=Betebenner&ff1=autBetebenner%2c+Damian+W.&id=EJ1054129

WestEd. (2014). *Kindergarten Readiness Assessment technical report fall 2014*. https://education.ohio.gov/getattachment/Topics/Early-Learning/Kindergarten/Ohios-Kindergarten-Readiness-Assessment/Kindergarten-Readiness-Assessment-for-Data-Manager/KRA_Technical_Report_2014_Final.pdf.aspx