ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-625

The Other Half of the Story: Does Excluding the Early Grades from School Ratings Matter?

Walter Herring
Mathematica

Because high-stakes testing for school accountability does not begin until third grade, accountability ratings for elementary schools do not directly measure students' academic progress in grades K through 2. While it is possible that children's test scores in grades 3 and above are highly correlated with children's outcomes in the untested grades, research provides reasons to believe that this might not be the case in all schools. This study explores whether measures of school quality based on test scores in grades 3 through 5 serve as a strong proxy for children's academic outcomes in grades K through 2. The results show that directly accounting for children's test scores in the early grades could lead to meaningful changes in schools' test-based performance ratings. The findings have important implications for accountability policy.

VERSION: August 2022

Suggested citation: Herring, Walter. (2022). The Other Half of the Story: Does Excluding the Early Grades from School Ratings Matter?. (EdWorkingPaper: 22-625). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/69qh-g416

The Other Half of the Story:

Does Excluding the Early Grades from School Ratings Matter?

Walter Herring

Mathematica

Abstract: Because high-stakes testing for school accountability does not begin until third grade, accountability ratings for elementary schools do not directly measure students' academic progress in grades K through 2. While it is possible that children's test scores in grades 3 and above are highly correlated with children's outcomes in the untested grades, research provides reasons to believe that this might not be the case in all schools. This study explores whether measures of school quality based on test scores in grades 3 through 5 serve as a strong proxy for children's academic outcomes in grades K through 2. The results show that directly accounting for children's test scores in the early grades could lead to meaningful changes in schools' test-based performance ratings. The findings have important implications for accountability policy.

Disclaimer: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200005 to the University of Virginia. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

I am grateful to NWEA for providing the data for this research through the Kingsbury Research Award.

Introduction

Since the advent of the education accountability movement in the 1980s and 1990s, policymakers and researchers have increasingly assessed school quality using standardized test scores (Jennings & Lauen, 2016). Under the Every Student Succeeds Act (ESSA), states are required to administer standardized assessments in math and reading to students beginning in third grade and hold schools accountable for children's performance on these assessments. While states also incorporate other factors like student attendance in schools' accountability ratings, roughly two-thirds of a school's overall rating is based on students' test scores on average across states' ESSA plans (Education Commission of the States, 2018; author's calculations). This test-based portion of school ratings is usually derived from a weighted combination of students' "achievement" levels (e.g., proficiency rates) and the amount of growth that students demonstrate on these assessments.

It is well documented that test-based ratings do not address all dimensions of school quality. By placing a heavy emphasis on children's scores on math and reading assessments, these measures do not account for non-academic factors like school safety (Bae, 2018) or untested subjects like science and social studies (Jacob, 2005; Winters et al., 2010). Comparatively little consideration has been given to the fact that, even in the narrow domains of math and reading, test-based measures of school quality do not directly account for students' progress in these subjects at all grade levels. This concern is particularly relevant in elementary schools in which half of students (those in grades 3 through 5) are required to take standardized assessments and half (those in grades K through 2) are not. Over the last decade, education policymakers have shown increasing interest in monitoring children's learning in the early

elementary grades (grades K through 2) through initiatives like kindergarten entry assessments (KEAs; Regenstein et al., 2017b). However, these early grades are largely excluded from test-based school quality ratings due to concerns associated with delivering high-stakes assessments to young children (e.g., Shephard, 1994; Solley, 2007).

While parents and policymakers might assume that elementary schools that are considered "high performing" based on math and reading test scores in grades 3 through 5 are also "high performing" in the untested early grades, research provides a number of reasons to suspect that this might not be the case. Fuller and Ladd (2013), for instance, show that schools that serve relatively small proportions of students that qualify for free or reduced-price lunch (FRPL) tend to staff more qualified teachers in the upper elementary grades than the lower elementary grades. This might suggest that current test-based quality ratings overestimate the academic progress made by children in grades K through 2 who attend schools that engage in this sort of strategic staffing. This finding also implies that certain schools (in this case, those serving fewer students that qualify for FRPL) might be particularly likely to receive ratings that overstate their contributions across all grade levels that they serve.

Understanding whether measures of school quality in grades 3 and above reflect the academic outcomes of children in the early elementary grades has important implications for education policy. If current ratings are not representative of children's learning in grades K through 2, then they would send an inaccurate (or at least incomplete) signal to policymakers and parents regarding how well schools support students' math and reading development across all grade levels. The consequences of this would be especially important if certain schools (e.g., those serving students from economically disadvantaged backgrounds) are particularly likely to be overrated or underrated based only on their test scores in grades three and above. However, no

study that I know of has directly addressed these issues. The dearth of research on this topic is largely due to the fact that researchers usually do not have access to the systemwide standardized assessment information needed to compare students' test scores in the early elementary grades to scores in the upper elementary grades precisely because the early grades are not subject to standardized tests under accountability policies.

Present Study

This study addresses this gap by using results from the widely-administered MAP Growth assessment. These data contain students math and reading test scores in grades K through 5 and allow me to answer the following research questions (RQs):

RQ1: To what extent are schools' achievement and growth scores in grades 3 through 5 correlated with their achievement and growth scores in grades K through 2? Does this relationship vary by school characteristics?

RQ2: How would the relative ranking of schools in a state change if schools' achievement and growth scores included test scores in grades K through 2? Are changes in rankings associated with school characteristics?

In answering the first research question, this study provides new evidence of the extent to which test-based school quality measures in the upper elementary grades that are currently subject to standardized tests differ from these same measures in the early elementary grades. If I observe systematic differences in achievement or growth scores across these grade bands, this would suggest that current test-based school ratings do not fully reflect the math and reading progress of all children who attend a given elementary school.

Even if there are differences in schools' achievement and growth scores in the upper and early elementary grades, these differences may not be sufficiently large in magnitude to be

meaningful in practice. Policymakers often make resource allocation decisions by separating schools into discreet groups based on how their accountability scores compare to other schools in the state. For example, ESSA requires policymakers to intervene in Title I schools that receive scores that fall in the bottom 5% of the state. In RQ2, I address whether ranking elementary schools based on students' test scores across all grade levels (K through 5) produces meaningfully different rankings than the status quo system that measures performance based only on scores in the upper elementary grades. Apart from overall differences, this question will also assess equity concerns by exploring whether schools serving different student populations are especially likely to benefit from (or are disadvantaged by) the status quo system which focuses only on grades 3 and up.

Literature Review

Measuring School Quality Using Test Scores

Educational accountability systems are intended to provide parents and policymakers with objective information regarding school performance. Parents make decisions regarding where to send their child to school based on performance ratings (Figlio & Ladd, 2015; Hastings & Weinstein, 2008; Imberman & Lovenheim, 2016). Policymakers use school performance metrics to reward high-achieving schools and intervene in schools that are deemed ineffective (US Department of Education, 2019). Standardized assessments are a critical component of accountability systems because, in theory, they provide a common measure of the academic progress that students are making in a given school (Figlio & Ladd, 2015). For decades, federal policy has required states to assign ratings to schools based heavily on standardized test results in math and reading in grades 3 through 8.

Under ESSA, the test-based portion of a school's rating is generally based on two distinct components: "achievement" and "growth." The achievement component is often measured by either the proportion of students in a school that meet state proficiency standards on the math and reading assessments in a given year or the average test scores among students at the school. Achievement measures like proficiency rates have been criticized because they tend to favor schools serving more advantaged student populations (Harris & Liu, 2018; McEachin & Polikoff, 2012). To address this shortcoming, policymakers in nearly all states now include some measure of academic growth in their school rating system, ranging from relatively simple "gain scores" to more complex value-added formulas (Data Quality Campaign, 2019).

States apply different weights to schools' achievement and growth scores when producing accountability ratings. On one extreme, 25% of school ratings in Kansas are determined by achievement ratings with no consideration given to growth (Kansas State Department of Education, 2018). On the other extreme, 54% of elementary school ratings in Mississippi are based on student growth on math and reading exams while 28% are based on achievement (Mississippi Department of Education, 2019). On average, though, states weigh achievement and growth roughly evenly, with each accounting for about 33% of a school's overall rating (Education Commission of the States, 2018; author's calculations). The remaining one-third of a school's rating is based on other metrics that vary across states, such as student attendance or the progress made by English learners (ELs) in the school.

Limitations of Test-Based Measures of School Quality

Of course, school ratings that are based largely on students' test scores in math and reading in grades 3 and beyond do not address all dimensions of school quality that policymakers value. Some measures, like school climate or students' social-emotional learning, are not

included because they are challenging to assess reliably (e.g., Hough et al., 2017). Other measures, like students' achievement in social studies and science, are excluded because of the costs associated with developing and administering more standardized assessments, as well as the lost instructional time that would result from an increase in testing. As Figlio and Ladd (2015) discuss, the emphasis on math and reading results thus represents policymakers' prioritization of children's progress in these subjects over other indicators of quality.

Even within the domains of math and reading, however, standardized assessments are not mandated until third grade. The exclusion of grades K through 2 from school accountability ratings is notable because education policymakers are increasingly interested in tracking children's academic progress in the early elementary grades (e.g., Mead, 2016; Petrilli, 2022). For example, many states now require that children take KEAs to assess their skills at kindergarten entry and use results from these assessments to inform decision making in elementary schools (Education Commission of the States, 2020; Regenstein et al., 2017b). The omission of the early grades from test-based school ratings, then, seems to stem less from policymakers' belief that children's academic progress in the upper elementary grades is more important than their progress in the early grades and more from the practical and political challenges associated with administering high-stakes assessments to young children (Regenstein et al., 2017b; Shephard, 1994; Solley, 2007). As a result, when making decisions based on school accountability ratings, parents and policymakers may make the implicit assumption that children's math and reading progress in third grade and beyond in a given school serves as a proxy for students' learning in these domains in the untested early grades as well.

Existing research, however, suggests that test-based measures of school quality in the upper elementary grades may not reflect the academic outcomes of students in the early

elementary years. First, by incentivizing schools to improve test scores in math and reading in third grade and beyond, accountability policies may contribute to differences in children's progress in these subjects in the tested and untested grades. School leaders have been shown to respond to accountability pressure by moving less-qualified and ineffective teachers from the upper elementary grades to early elementary classrooms (Fuller & Ladd, 2013; Grissom et al., 2017). Grissom et al.'s (2017) analysis reveals that moving an ineffective teacher from a tested classroom to an untested classroom has a negative impact on the low-stakes assessment scores of the teacher's new students. Adding to this body of evidence, qualitative work conducted by Diamond and Spillane (2004) shows that schools devote more instructional and professional development resources to the upper elementary grades than the lower elementary grades. Taken together, these behaviors could indicate that schools' test-based ratings under accountability systems may overstate the progress that children make in untested grades in schools that employ strategic strategies.

Apart from the unintended consequences of accountability policies, there are other reasons to suspect that test-based measures of school quality in grades 3 and above might not reflect children's progress in math and reading in the early grades. For example, Cameron et al. (2015) show that children learn at the fastest rate prior to third grade, indicating that schools may exhibit more growth in the early elementary grades relative to the later grades. These growth rates also differ by racial and socioeconomic subgroups: Disparities in math and reading outcomes between children from higher- and lower-income households, for instance, grow during the early elementary grades. This finding could indicate that schools serving larger populations of children from economically disadvantaged backgrounds might exhibit less growth in the early elementary grades relative to the upper elementary grades. Reardon et al. (2015) also

show that disparities in math skills between Hispanic children and their White peers shrink during the early grades, suggesting that schools serving large populations of Hispanic children may exhibit more progress in math in the early grades compared to other schools.

Taken together, this evidence suggests that schools' achievement and growth measures based on scores in grades 3 and above may not proxy children's progress in math and reading in grades K through 2 in some schools. Further, they demonstrate that mismatches between test-based measures of school quality in grades 3 through 5 and students' academic outcomes in the untested grades might be especially likely in schools serving particular student populations (e.g., students from economically disadvantaged backgrounds). To date, however, no study that I know of has directly assessed whether students' test scores in the early and upper elementary grades systematically differ or explored the implications that such differences carry for the inferences drawn based on school ratings.

Potential Consequences of Differences in Test-Based Ratings Across Grades

Understanding the extent to which schools' test-based ratings in the upper elementary grades differ from their ratings in the early elementary grades has important implications for the decisions that policymakers make based on school ratings. Parents, for example, have been shown to change their school-choice behaviors when presented with test-based measures of school quality (Hastings & Weinstein, 2008; Imberman & Lovenheim, 2016). If test-based ratings of certain schools based on grades 3 through 5 systematically misrepresent the progress that students make in the untested grades, parents would decide where to send their child to school based on incomplete, and perhaps misleading, information. Differences in test scores across grade bands would also impact resource allocation decisions made based on school ratings: If schools' test-based ratings in grades 3 through 5 are not a strong proxy for their

performance in the early grades, this would imply that, by focusing only on grades 3 and above, policymakers might misclassify certain schools as in need of support under accountability regimes or fail to identify schools that would benefit from additional resources (Harris & Liu, 2018).

In order to assess whether schools' achievement and growth scores in the upper elementary grades reflect the progress that children make in math and reading in the untested grades, this study uses unique data containing children's standardized assessment scores in each grade between kindergarten and fifth grade. The results provide new evidence regarding the extent to which elementary schools' achievement and growth scores in grade 3 through 5 reflect students' achievement and growth in the untested grades. This paper also addresses how differences in test scores across grade bands could change the decisions parents and policymakers make based on school ratings and how these differences correlate with school characteristics.

Data

I use results from the MAP Growth assessment administered by the research non-profit NWEA to calculate test-based measures of school quality. NWEA partners with more than 24,000 public schools across all 50 states and the District of Columbia to deliver the MAP Growth assessment (Thum & Kuhfeld, 2020). My data contain millions of test events across a five-year panel (2013-14 to 2018-19) in ten states [1].

Math and Reading Test Scores

Children's math and reading skills were measured in the spring of each school year between 2013 and 2018 using the MAP Growth assessment. MAP Growth is a computer-adaptive assessment that asks children questions based on their estimated skill level so that

children are not exposed to content that is either too easy or too difficult for them. Children's raw scores on the assessment are transformed to RIT (Rasch Unit) Scale scores based on item difficulty. These RIT scores range from 100 to 350 (Thum & Kuhfeld, 2020). The assessment is both vertically aligned and interval scaled, making it well suited for calculating growth-based estimates of school performance (e.g., Soland, 2019).

School-Level Covariates

The data include school-level information provided by NWEA. These supplementary school-level data were linked by NWEA staff from the Common Core of Data (CCD) prior to anonymizing school identifiers. They include the total number of students enrolled in the school, the percentage of students in each racial and ethnic subgroup, the percentage of students who qualify for FRPL, and the locale of the school (city, rural, suburban, or town).

Sample Description

I exclude schools that administered MAP Growth to 5 or fewer children in any grade between kindergarten and fifth grade and schools that had far fewer test-takers in the early elementary grades relative to the upper elementary grades [2]. Of the roughly 11,000 schools in my sample, about 1,100 were excluded because they did not offer MAP Growth to a sufficient number of students in each grade, 50 were excluded because they offered the assessment to far fewer children in the early grades compared to the upper elementary grades, and 4,400 were excluded because they met neither of these two conditions.

As shown in the first two columns of Table 1, schools that meet these criteria differ from schools that do not in several respects. Schools that remain in my sample had larger student populations and tended to serve relatively smaller proportions of Hispanic children and White

children and larger proportions of Black children and children that qualified for FRPL. They were also more likely to be located in cities or rural areas.

- TABLE 1 ABOUT HERE -

It is also important to consider that this sample is not representative of schools nationwide or schools within each of the deidentified states included in my data. Leaders at a given school or school district must partner with NWEA to offer the assessment. To understand the implications of the "opt-in" nature of MAP Growth, I compare the characteristics of schools in my sample to those of the universe of schools in CCD that enrolled kindergarten students in the 2017-18 school year in column (3) of Table 1. Schools in my sample served smaller proportions of Asian, Hispanic, Native American, and Pacific Islander students and larger proportions of Black students and students who qualify for FRPL than schools nationwide. Schools in my sample were less likely to be located in a suburb and more likely to be located in a city than elementary schools nationwide. The implications of these differences as they relate to the external validity of my findings are discussed in the limitations section.

Methodology

Calculating Achievement and Growth Scores

In this paper, I construct test-based measures of school quality that are similar to those used by states in their accountability systems. I mirror the approach taken in most states and calculate separate achievement and growth scores for each school in my sample. In order to assign schools' achievement scores, I compute the average MAP Growth scores of students attending each school in a given year. While the majority of states measure academic achievement by calculating the proportion of students that meet state proficiency standards in math and reading in a given school, I calculate average test scores here because there are no pre-

established thresholds for proficiency associated with MAP Growth. Measuring schools' achievement score by calculating average test scores is an alternative approach used in states like Colorado and Connecticut. Because the MAP Growth data are vertically scaled, children in the early grades have scores that are systematically lower than children in higher grades. In order to ensure that comparisons of average scores across grade levels are made on the same scale, I standardize scores within state, grade level, subject, and year.

There are numerous metrics that different states use to assign growth scores to schools under accountability systems. In this study, I employ the growth measure most commonly used by states in their accountability system: the Student Growth Percentile (SGP; Betebenner, 2011). SGPs use quantile regression techniques to assess the growth that a student exhibited in a given year relative to their peers who have a similar history of test scores. For example, a student with an SGP of 75 experienced more growth in math or reading in a given school year than 75 percent of students in the same grade and state who received the same score(s) on their prior assessment(s) [3]. Following the approach taken by states like Colorado and Maryland, I assign growth scores to schools based on the median SGP among children who attended that school.

Computing average test scores and median SGPs is not the most methodologically robust way to measure school quality. Indeed, both of these measures have been criticized because they do not account for student demographic characteristic, leading to concerns that schools serving historically disadvantaged student populations might score systematically lower on these measures (McEachin & Polikoff, 2012; Walsh & Isenberg, 2015). This paper, however, is intended to assess the extent to which public perception of school quality, as measured through test-based accountability metrics, might change if the early elementary grades were incorporated into schools' accountability ratings. In this respect, average scores and SGPs are appropriate

measures of school quality because they are among the most commonly used test-based measures across states' accountability plans (Data Quality Campaign, 2019).

RQ1: To what extent are schools' achievement and growth scores in grades 3 through 5 correlated with their achievement and growth scores in grades K through 2? Does this relationship vary by school characteristics?

To compare schools' achievement scores in the upper and early elementary grades, I calculate a school's average test scores in grades K through 2 which are not typically subject to standardized assessments ($\mu_{untested}$) and their average scores in grades 3 through 5 in which standardized assessments are mandated (μ_{tested}). Note that $\mu_{untested}$ and μ_{tested} represent school-by-year measures such that most schools in the data will have multiple measures of each parameter. Because states assign scores to schools based on both math and reading test results, I average schools' reading score and math score to produce $\mu_{untested}$ and μ_{tested}. Results are very similar when I compare scores for reading and math separately (see Appendix B).

Borrowing from an approach taken by McEachin and Atteberry (2017), I compare mean test scores across grade levels by regressing $\mu_{untested}$ on μ_{tested} using ordinary least squares (OLS) regression to determine the strength of the relationship between the two achievement ratings. This regression takes the form:

$$(1)\ \ \mu_{untested} = \alpha + \beta\mu_{tested} + \varepsilon$$

Where β represents the coefficient of interest and reflects the relationship between a school's average test score in the untested and tested grades. To account for the correlation between an individual school's achievement rating across multiple years, I cluster standard errors at the school level. I also plot $\mu_{untested}$ against μ_{tested} in a scatterplot and compare these results to the 45-degree line to provide a visual representation of this relationship in Appendix C.

I next consider the extent to which school characteristics are related to differences in their achievement and growth scores in the untested early grades relative to the tested grades. To do so, I run a series of OLS regressions similar to equation (1):

$$(2)\ \mu_{untested} = \alpha + \beta\mu_{tested} + \boldsymbol{\pi X} + \varepsilon$$

Where $\boldsymbol{X}$ represents one or a series of school-level covariates, including the percentage of students qualifying for FRPL and the percentage of children in each racial or ethnic subgroup in the school. The coefficient(s) $\boldsymbol{\pi}$ reflect the extent to which each of these covariates is related to a school's achievement score in the untested grades when holding the school's scores in the tested grades and any other covariates included in $\boldsymbol{X}$ constant. Standard errors for these regressions are clustered at the school level.

I conduct a similar analysis for schools' growth scores. I assign growth scores to schools separately based on their median SGP in grades K through 3 ($\Delta_{untested}$) and their median SGP in grades 3 through 5 (Δ_{tested}) in a given year [4]. I then compare $\Delta_{untested}$ and Δ_{tested} using the same regression and graphical approaches described above.

RQ2: How would the relative ranking of schools in a state change if schools' achievement and growth scores included test scores in grades K through 2? Are changes in rankings associated with school characteristics?

The second research question assesses the extent to which public perception of school quality might change if their achievement and growth scores were based on students' test results in grades K through 5. To do so, I average schools' achievement and growth scores in order to produce a "combined" test-based performance score for each school. While the preceding questions consider achievement and growth scores separately, states generally assign ratings to schools based largely on a weighted combination of these two components. I transform each

school's achievement and growth scores in the upper elementary grades (μ_{tested} and Δ_{tested}, respectively) into percentile scores, ranging from 0 to 100, relative to other schools in the same state and year. I then average these two percentile scores together to produce the "combined" score. In this way, a school's achievement and growth scores are weighted evenly to produce the combined score, which ranges from 0 to 100. In the main body of this text, I choose to weight achievement and growth equally because, on average, states weight these two components evenly in their accountability systems (Education Commission of the States, 2018; author's calculations). In Appendix D, I present results when alternative weights are employed.

To compare schools' combined achievement and growth scores in third through fifth grade to the scores they would have received if they included students' test results from the early elementary grades, I calculate achievement and growth scores based on children's MAP Growth results in grades K through 5 (μ_{all} and Δ_{all}, respectively) to produce a separate combined score for each school. Next, I rank schools separately based on their combined achievement and growth scores across these two different grade bands (3-5 and K-5) within each state and year.

With schools ranked in this manner, I then assess the difference in schools' relative rankings across the two grade bands in two ways. First, following the approaches taken by McEachin and Atteberry (2017) and Harris and Liu (2018), I divide schools into five quintiles based on their rank in grades 3 through 5 and their rank in grades K through 5 in order to construct "transition matrices" that reflect the proportion of schools that change quintiles depending on the grade levels included in the achievement and growth score calculations. Second, based on ESSA's provision that states intervene in Title I schools with performance ratings in the bottom 5%, I construct a similar transition matrix reflecting the proportion of schools scoring above and below the bottom 5% threshold in each state based on scores in grades

3 through 5 and those scoring above and below the threshold in grades K through 5. Because some states make high-stakes decisions about schools based on multi-year averages of their performance scores rather than results from a single year, I report these same transition matrices derived from three-year averages of schools' scores in Appendix E with very similar results.

The results from these analyses are difficult to interpret on their own because schools might move between quintiles or across the bottom 5% threshold for a number of reasons that are not related to their "true" achievement and growth scores. Changes could be driven by smaller schools, which have been shown to have particularly volatile achievement and growth scores (Kane & Staiger, 2002), or by the statistical "noise" introduced by adding children from grades K through 2 to achievement and growth calculations. In Appendix F, I address these concerns and show that my findings do not appear to be driven by either of these phenomena.

Even if relatively few schools move between quintiles or across the bottom 5% threshold, any differences in schools' relative ranking could nonetheless have important consequences for educational equity if schools serving particular student populations are more likely to see their ranking decrease or increase after accounting for students' test scores in grades K-2. I compare the characteristics of schools that increased their ranking to those that saw their ranking decrease by employing an OLS regression of the form:

$$(3)\ Chacteristic_{sy} = \alpha + \gamma DecreasedRank_{sy} + \varepsilon_{sy}$$

Where $Chacteristic_{sy}$ represents a characteristic of school *s* in year *y* (e.g., the proportion of children qualifying for FRPL or whether the school is located in a rural area). $DecreasedRank_{sy}$ is a binary variable indicating whether school *s* in year *y* decreased quintiles or moved above the bottom 5% threshold after incorporating children's test scores in grades K through 2 in achievement and growth calculations. Because I restrict the analysis to schools that

changed quintiles or bottom 5% threshold status after accounting for grades K through 2, the coefficient γ indicates the extent to which schools that decreased quintiles or bottom 5% status differed from schools that saw their quintile or 5% status increase along the particular school characteristic. To account for the correlation between school-by-year observations of the same school, I again cluster standard errors at the school level.

Results

RQ1: To what extent are schools' achievement and growth scores in grades 3 through 5 correlated with their achievement and growth scores in grades K through 2? Does this relationship vary by school characteristics?

Schools' achievement scores in grades 3 through 5 were highly correlated with their achievement scores in grades K through 2. As shown in column (1) of Table 2, the regression outlined in equation 1 yields a β coefficient of 0.877 for schools combined math and reading achievement scores. By contrast, Table 3 shows that schools' growth scores (as measured by the median SGP) in the upper elementary grades tend to be very different than their scores in the lower elementary grades. Regressing median SGPs in grades K through 2 on those in grades 3 through 5 as in equation (1) yields a coefficient of 0.40, reflecting a much weaker relationship between SGP measures across grade levels.

- TABLE 2 ABOUT HERE –

- TABLE 3 ABOUT HERE -

I next explore how these differences in achievement and growth across grade bands relate to school characteristics. Column (2) of Table 2 shows that, holding schools' average scores in the tested grades constant, schools serving higher proportions of students that qualified for FRPL tended to have lower average test scores in the early elementary grades: A ten percentage point

increase in the proportion of children qualifying for FRPL in a school is associated with a one-tenth of a standard deviation decrease in average test scores in grades K through 2, holding the school's achievement score in the upper elementary grades constant. This trend holds in column (3) when I control for the racial and ethnic backgrounds of students who attend the schools as the coefficient on the "% FRPL" covariate remains unchanged.

In Table 3, I present similar results exploring the relationship between school characteristics and schools' growth scores in the untested grades. The results for schools' growth measures are qualitatively similar to those reported in Table 2. The results in column (2) suggest that schools serving larger proportions of students that qualify for FRPL had lower growth scores in the untested early grades after controlling for their growth scores in the tested grades. Holding schools' median SGP in the tested grades constant, a ten percentage point increase in the proportion of children who qualify for FRPL in a school is associated with a one point decrease in the school's median SGP in the untested grades. This trend also holds after controlling for the proportion of children in each racial and ethnic subgroup that attend the school, though the coefficient on the FRPL term decreases slightly.

RQ2: How would the relative ranking of schools in a state change if schools' achievement and growth scores included test scores in grades K through 2? Are changes in rankings associated with school characteristics?

In answering this research question, I explore the potential consequences of the differences in achievement and growth ratings across different grade levels documented in RQ1. Table 4 below displays a transition matrix comparing schools' within-state-and-year quintile ranking based on their combined achievement and growth scores in grades 3 through 5 against their rankings after incorporating scores for all students in grades K through 5. The results show

that a large percentage of schools would be ranked in a different quintile if their test-based performance scores included children's test results in grades K through 2: 42% of schools change quintiles after accounting for students test scores in grades K-2, with 5% moving multiple quintiles.

- TABLE 4 ABOUT HERE -

Table 5 displays a transition matrix reflecting the proportion of schools with scores above or below the bottom 5% threshold across the different grade bands. Roughly 38% of schools that fall in the bottom 5% based on grades 3 through 5 no longer fall in the bottom 5% when grades K through 2 are accounted for.

- TABLE 5 ABOUT HERE -

I next consider how the movement across quintiles or the bottom 5% threshold relates to school characteristics. Table 6 below documents the characteristics of schools that increase or decrease quintiles after including results in grades K through 2. The results reveal that schools that decreased quintiles served larger proportions of children who qualified for FRPL and Black children and smaller proportions of White children than schools that increased quintiles. Schools that decreased quintiles were also more likely to be located in cities and less likely to be located in suburbs.

- TABLE 6 ABOUT HERE -

I last explore the characteristics of schools that move across the bottom 5% threshold after accounting for children's MAP Growth scores in the early elementary grades. Table 7 shows that schools that fell below the bottom 5% threshold after including early elementary scores served much larger populations of Black children and smaller populations of Native

American and White children than the schools that replaced them in the bottom 5% on average. Schools that fell below the threshold were also more likely to be located in cities and less likely to be located in rural areas.

- TABLE 7 ABOUT HERE -

Discussion

This paper provides evidence that measures of school quality based on students' math and reading scores in grades 3 through 5 often do not reflect the math and reading outcomes of children in the untested early elementary grades. This misalignment could have important consequences for schools: Many schools that are labeled as particularly low performing based on students' test scores in grades 3 through 5 would not be labeled as such if ratings incorporated children's academic achievement and growth in the early elementary grades. Conversely, schools serving larger populations of Black children and children from economically disadvantaged backgrounds were particularly likely to see their test-based rating decrease after including test results from grades K through 2. Below I discuss the limitations of this analysis, as well as the implications of my findings for policy and future research.

Limitations

Though these results have important implications for education policy, this study carries a number of limitations. First, while thousands of schools in the ten states included in my data offered MAP Growth in each grade between kindergarten and fifth grade, the sample of schools included here is not representative of the universe of schools nationally. In particular, because they offer MAP Growth in the early grades, schools in my sample might devote more resources to children's learning in grades K through 2 than schools that do not offer the assessment in these years. This could mean that schools in my sample are especially likely to show differences in

students' math and reading progress in the untested and tested grade levels compared to schools not included. As such, the results I report here may not reflect how schools' achievement and growth scores differ across grade bands for schools nationwide. Future work ought to explore whether these findings would be similar if a representative sample of schools were employed.

Second, I calculate schools' test-based ratings using results from a low-stakes assessment that is not tied to accountability policy. Because there are generally no consequences attached to students' or schools' performance on MAP Growth, it is possible that students and teachers may not exert as much effort on MAP Growth as they do on the high-stakes assessments used by states to calculate accountability scores. Further, teachers are less likely to "teach to the test" on a low-stakes assessment like MAP Growth than they would be on a higher-stakes assessment. To the extent that the specific skills covered on state assessments and the skills covered on MAP Growth do not overlap, it might be that the test-based measures of school quality that I calculate here would underestimate the ratings that schools would receive in the upper elementary grades were I to use results from a high-stakes assessment (Jacob, 2017; Jennings & Bearak, 2014). As such, my findings may have been different if a high-stakes assessment were used. It should be noted, however, that NWEA takes care to align MAP Growth with state math and reading standards, which may ameliorate these concerns.

Last, the school ratings I calculate in this paper are limited in important respects. While math and reading scores account for the majority of a school's rating under accountability, other factors like chronic absenteeism and school climate are also incorporated into school ratings in practice (Education Commission of the States, 2018). My data only contain children's math and reading test scores, and so I cannot include these non-test indicators of school quality in the school ratings I produce here. Additionally, because the states in my data were deidentified by

NWEA prior to being shared with me, I cannot mimic the scoring system (e.g., the particular growth measure used or the weights applied to achievement and growth scores) employed in each state and instead apply the same approach across all ten states. As a result of these limitations, the school performance ratings I report here are at best a rough approximation of the actual ratings a school would receive in their accountability system in practice.

Implications

With these limitations in mind, the results of this paper reveal important takeaways for education policymakers. The paper provides evidence that schools' achievement levels in third through fifth grade on average serve as a strong proxy for their achievement levels in the early elementary grades. On the other hand, schools' median SGPs in the early grades tended to be very different than their median SGPs in the upper elementary grades, suggesting that current growth ratings based on students' scores in grades 3 through 5 often do not reflect the academic growth experienced by children in the early elementary grades. These findings are consistent with prior work which shows that test score levels are relatively stable year to year while growth scores vary considerably (Kane & Staiger, 2002).

My results also imply that the differences in achievement and growth scores across grade levels could translate to meaningful differences in the decisions that parents and policymakers make based on these scores. Incorporating children's math and reading scores between kindergarten and second grade in school ratings yields considerably different rankings than rankings which only use results from grades that are currently subject to standardized tests. This is to say that, under the status quo system, many schools' ratings understate their performance across all grade bands while others' overstate their performance.

The movement of schools across quintile and bottom 5% thresholds documented in this study is especially consequential because it is strongly associated with school characteristics. Schools that saw their quintile ranking decrease after incorporating MAP Growth scores in the early grades served larger proportions of children who qualified for FRPL and Black children than schools that increased their quintile ranking (Table 6), reflecting the fact that schools serving more children who qualify for FRPL had lower achievement and growth scores in the early elementary grades than those serving more advantaged student populations (Tables 2 and 3). The relationships between student characteristics and changes in school rankings appear consistent with prior work which suggests that test score disparities along lines of both race and socioeconomic status tend to increase during the early elementary grades (Cameron et al., 2015).

The correlation between school rating changes and student demographics raises important considerations for educational equity, and how one interprets these results largely depends on how one views the purpose and consequences of accountability policy. On the one hand, one of the purported aims of measuring school quality is to help policymakers identify struggling schools so that they can provide them with additional resources and implement interventions to improve students' outcomes. Consistent with this assertion, existing evidence suggests that students' test scores do improve in schools that have been labeled as low-performing (Saw et al., 2017; Winters & Cowen, 2012). Seen through this lens, the findings in this study suggest that, by measuring school quality using test results in third grade and beyond, states might fail to intervene in many schools serving comparatively large proportions of Black children and children who qualify for FRPL that would benefit from these supports based on students' test scores in the early elementary grades. From this perspective, moving to an accountability system

that incorporates children's learning in grades K through 2 could promote equity by directing more support and resources to schools serving these student populations.

On the other hand, low ratings have a number of consequences beyond the supports and interventions prescribed by policy. Schools that receive low accountability scores struggle to retain teachers (Feng et al., 2010; Hanushek & Rivkin, 2010), and parents are less likely to send their children to schools with low test scores (Hastings & Weinstein, 2008). Further, low accountability ratings can hamper schools' ability to secure donations (Figlio & Kenny, 2009) and may even impact housing prices in the school's neighborhood, at least in the short term (Figlio & Lucas, 2004). In light of this evidence, an alternative interpretation of this research is that moving to a system that incorporates students' learning in the early elementary years would serve only to disadvantage schools that serve larger populations of Black children and children who qualify for FRPL as their ratings tend to decrease under such a system.

These questions are difficult to grapple with, but they are critical to confront as policymakers consider how they will measure school quality under ESSA. This paper provides evidence that current test-based ratings of schools do not adequately reflect the learning outcomes of children in the untested grades in many cases. Further, by overemphasizing test scores in the upper elementary grades, current accountability systems provide practitioners with strong incentives to engage in practices that could prove detrimental to children's learning in the early elementary grades (Diamond & Spillane, 2004; Fuller & Ladd, 2013; Grissom et al., 2017). Given this evidence, policymakers ought to consider means by which they could produce school ratings that better reflect children's learning across all grade levels. Doing so, however, will necessitate further inquiry. Administering high-stakes assessments to children in the early years is controversial (e.g., Solley, 2007), and existing assessments administered in the early grades,

like KEAs, are not designed to be used in accountability systems (Dragoset et al., 2019; Regenstein et al., 2017b). As such, policymakers would have to turn to alternative measures of school performance like attendance or observations in order to incorporate indicators of children's learning in the early elementary grades in school ratings (Aldeman, 2016; Regenstein et al., 2017a). Before moving in this direction, though, it will be critical to validate these alternative indicators and explore the extent to which they are related to children's academic achievement and growth in the early grades.

Regardless of how children's early academic outcomes are incorporated into school ratings under accountability, the findings in this paper speak to the need to provide increased visibility to children's learning in the early elementary grades. Future work, both quantitative and qualitative, ought to assess why some schools make greater or lesser contributions in the untested grades relative to grades that are subject to standardized tests and explore means by which we might reverse the troubling trends documented in this paper. Answering questions like these requires systematic information about children's learning in the early grades. In this respect, policymakers should continue to make use of data from KEAs and other assessments to track children's learning in the early years and identify schools which may need additional resources to support their youngest learners (Regenstein et al., 2017b).

Endnotes:

[1] For privacy reasons, all student, school, district, and state identifiers have been removed and replaced with NWEA-specific identifiers. As such, I am able to group students into schools, districts, and states, but I am unable to identify which specific school, district, or state each unit

represents. The ten states were selected because they offered MAP Growth in the greatest number of elementary schools.

[2] To do so, I tally the total number of test takers in a given school and year in grades K through 2 and the total number of test takers in grades 3 through 5. If the ratio of test takers across grade bands is less than 0.5 or greater than 1.5, I remove the school from the sample.

[3] For a discussion of the benefits and limitations of SGPs, see Appendix A.

[4] Under the status quo system, third grade is usually considered a "baseline" value on which growth cannot be calculated. As such, I include students' growth between second and third grade in the "untested" growth measure in this study since it is usually not accounted for in current school ratings.

Tables

Table 1: School Characteristics

	NWEA Data		CCD Data
	(1) In Sample	(2) Not in Sample	(3) All Schools in US
Number of Students	492.3	458.4	457.1
Sociodemographic			
% FRPL	61.0%	52.4%	56.0%
% Asian	3.5%	4.2%	4.7%
% Black	23.4%	14.3%	15.9%
% Hispanic	19.8%	21.3%	25.1%
% Native American	0.7%	0.7%	2.0%
% Pacific Islander	0.1%	0.1%	0.5%
% Two or More Race	4.0%	4.2%	4.5%
% White	48.4%	55.0%	49.9%
School Locale			
% City	38.3%	26.7%	30.4%
% Rural	23.4%	20.7%	26.3%
% Suburb	29.7%	37.7%	33.2%
% Town	8.7%	15.0%	10.2%
Number of Schools	5193	5569	52699

Note: CCD figures calculated from 2017-2018 school year. NWEA figures are averages across five year panel based on schools offering the MAP Growth math assessment. Results are very similar for schools offering the MAP Growth reading exam. “In Sample” denotes schools that administered the MAP Growth math assessment to at least 6 children in each grade between kindergarten and fifth grade and who administer the assessment to a similar number of children in the early and upper elementary grades.

Table 2: Associations Between School Characteristics and Achievement Scores in Math and Reading in the Untested Grades

	(1)		(2)		(3)	
Achievement Score Tested Grades	0.877	***	0.820	***	0.822	***
% FRPL			-0.001	***	-0.001	***
% Asian					0.003	
% Black					0.002	
% Hispanic					0.001	
% Native American					0.003	
% Pacific Islander					0.000	
% Two or More Race					0.003	
% White					0.002	
(Intercept)	0.013	***	0.088	***	-0.092	

n=17,092 school-by-year observations. Standard errors are clustered at the school level.
*p<.10 **p<.05 ***p<.01

Table 3: Associations Between School Characteristics and Growth Scores in the Untested Grades

	(1)		(2)		(3)	
Growth Score Tested Grades	0.397	***	0.306	***	0.280	***
% FRPL			-0.107	***	-0.074	***
% Asian					0.224	**
% Black					0.102	
% Hispanic					0.163	*
% Native American					0.144	
% Pacific Islander					-0.159	
% Two or More Race					0.150	
% White					0.160	
(Intercept)	29.493	***	40.399	***	24.784	**

n=11,967 school-by-year observations. Standard errors are clustered at the school level.
*p<.10 **p<.05 ***p<.01

Table 4: Transition Matrix of Combined Rating Quintiles Math and Reading

Quintile (Grade 3-5)	Quintile (Grade K-5)				
	1	**2**	**3**	**4**	**5**
1	15.10%	4.41%	0.63%	0.03%	0.00%
2	4.14%	9.64%	5.20%	1.06%	0.08%
3	0.84%	4.88%	8.52%	5.03%	0.75%
4	0.08%	1.14%	4.93%	9.38%	4.35%
5	0.00%	0.03%	0.75%	4.40%	14.62%

n=11,967 school-by-year observations

Table 5: Bottom 5% Matrix for Combined Scores Math and Reading

	Grades K-5	
Grades 3-5	**Not Bottom 5%**	**Bottom 5%**
Not Bottom 5 %	92.77%	2.01%
Bottom 5%	1.96%	3.26%

n=11,967 school-by-year observations

Table 6: Characteristics of Schools that Increase or Decrease Quintiles After Accounting for K-2 Scores – Math and Reading

	Increased	Decreased	Difference	
Total Test Takers	391.26	382.49	-8.76	
% FRPL	58.68%	62.75%	4.07	***
% Asian	3.48%	3.36%	-0.12	
% Black	18.13%	21.80%	3.67	***
% Hispanic	17.88%	18.43%	0.55	
% Native American	0.64%	0.51%	-0.13	
% Pacific Islander	0.08%	0.09%	0.01	
% Two or More Race	4.14%	4.22%	0.08	
% White	55.52%	51.40%	-4.11	***
% City	31.74%	35.35%	3.60	**
% Rural	27.78%	26.75%	-1.03	
% Suburb	32.21%	29.19%	-3.02	*
% Town	8.28%	8.72%	0.44	
N	2577	2537		

*p<.10 **p<0.05 ***p<0.01

Table 7: Characteristics of Schools That Move Across the Bottom 5% Threshold - Math and Reading

	Rise Above Threshold After K-2 Included	Fall Below Threshold After K-2 Included	Difference	
Total Test Takers	302.24	324.83	22.58	
% FRPL	83.67%	83.44%	-0.24	
% Asian	1.47%	1.69%	0.22	
% Black	40.56%	54.87%	14.31	***
% Hispanic	20.43%	17.79%	-2.64	
% Native American	2.65%	0.26%	-2.40	*
% Pacific Islander	0.06%	0.12%	0.06	
% Two or More Race	3.77%	4.02%	0.25	
% White	30.93%	21.12%	-9.82	***
% City	49.15%	65.15%	16.00	***
% Rural	25.21%	14.11%	-11.11	***
% Suburb	19.23%	14.11%	-5.12	
% Town	6.41%	6.64%	0.23	
N	234	241		

*p<.10 **p<0.05 ***p<0.01

Works Cited

Aldeman, C. (2016). *Grading Schools: How States Should Define "School Quality" Under the Every Student Succeeds Act*. Bellwether Education Partners. https://bellwethereducation.org/sites/default/files/Bellwether_GradingSchools_FINAL101916.pdf. Accessed March 8, 2022.

Bae, S. (2018). Redesigning systems of school accountability: A multiple measures approach to accountability and support. *Education Policy Analysis Archives*, *26*, 8. https://doi.org/10.14507/epaa.26.2920

Betebenner, D. W. (2011). *A Technical Overview of the Student Growth Percentile Methodology*. The National Center for Improvement of Educational Assessment. https://ksde.org/Portals/0/Research%20and%20Evaluation/SGP_Technical_Overview.pdf. Accessed June 4, 2021.

Cameron, C. E., Grimm, K. J., Steele, J. S., Castro-Schilo, L., & Grissmer, D. W. (2015). Nonlinear Gompertz curve models of achievement gaps in mathematics and reading. *Journal of Educational Psychology*, *107*(3), 789–804. https://doi.org/10.1037/edu0000009

Data Quality Campaign. (2019). *Growth Data: It Matters, and It's Complicated*. https://dataqualitycampaign.org/wp-content/uploads/2019/04/DQC-Growth-Data-Resources.pdf. Accessed June 2, 2021.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality. *The Teachers College Record*, 1145–1176.

Dragoset, L., Baxter, C., Dotter, D., & Walsh, E. (2019). Measuring School Performance for Early Elementary Grades in Maryland. In *Regional Educational Laboratory Mid-Atlantic*. Regional Educational Laboratory Mid-Atlantic. https://eric.ed.gov/?id=ED601956

Education Commission of the States. (2018). *Accountability and Reporting: Current System*. Education Commission of the States. https://ecs.secure.force.com/mbdata/mbQuest5E?rep=SA172. Accessed September 23, 2019.

Education Commission of the States. (2020). State K-3 Policies. Education Commission of the States. https://reports.ecs.org/comparisons/state-k-3-policies-05. Accessed July 12, 2022.

Feng, L., Figlio, D. N., & Sass, T. (2010). *School Accountability and Teacher Mobility* [NBER Working Paper]. https://www.nber.org/papers/w16070.pdf

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, *93*, 1069–1077. https://doi.org/10.1016/j.jpubeco.2009.07.003

Figlio, D. N., & Ladd, H. F. (2015). School Accountability and Student Achievement. In *Handbook of Research in Education Finance and Policy* (2nd ed.). Routledge.

Figlio, D. N., & Lucas, M. E. (2004). What's in a Grade? School Report Cards and the Housing Market. *American Economic Review*, *94*(3), 591–604. https://doi.org/10.1257/0002828041464489

Fuller, S. C., & Ladd, H. F. (2013). School-Based Accountability and the Distribution of Teacher Quality Across Grades in Elementary School. *Education Finance and Policy*, *8*(4), 528–559. https://doi.org/10.1162/EDFP_a_00112

Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic Staffing? How Performance Pressures Affect the Distribution of Teachers Within Schools and Resulting Student Achievement. *American Educational Research Journal*, *54*(6), 1079–1116. https://doi.org/10.3102/0002831217716301

Hanushek, E. A., & Rivkin, S. G. (2010). The Quality and Distribution of Teachers under the No Child Left Behind Act. *Journal of Economic Perspectives*, *24*(3), 133–150. https://doi.org/10.1257/jep.24.3.133

Harris, D. N., & Liu, L. (2018). *What Gets Measured Gets Done: Multiple Measures, Value-Added, and the Next Generation of Accountability under ESSA*. 41. https://educationresearchalliancenola.org/files/publications/051418-Harris-Liu-What-Gets-Measured-Gets-Done.pdf. Accessed March 9, 2022.

Hastings, J. S., & Weinstein, J. M. (2008). Information, School Choice, and Academic Achievement: Evidence from Two Experiments. *The Quarterly Journal of Economics*, *123*(4), 1373–1414. https://doi.org/10.1162/qjec.2008.123.4.1373

Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using Surveys of Students' Social-Emotional Learning and School Climate for Accountability and Continuous Improvement* (p. 38). Policy Analysis for California Education. https://edpolicyinca.org/publications/using-surveys-students-social-emotional-skills-and-school-climate-accountability. Accessed March 8, 2022.

Imberman, S. A., & Lovenheim, M. F. (2016). Does the market value value-added? Evidence from housing prices after a public release of school and teacher value-added. *Journal of Urban Economics*, *91*, 104–121. https://doi.org/10.1016/j.jue.2015.06.001

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. (2017). The Changing Federal Role in School Accountability: Point/Counterpoint. *Journal of Policy Analysis and Management*, *36*(2), 469–477. https://doi.org/10.1002/pam.21975

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the Test" in the NCLB Era: How Test Predictability Affects Our Understanding of Student Performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Jennings, J. L., & Lauen, D. L. (2016). Accountability, Inequality, and Achievement: The Effects of the No Child Left Behind Act on Multiple Measures of Student Learning. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *2*(5), 220–241. https://doi.org/10.7758/RSF.2016.2.5.11

Kane, T. J., & Staiger, D. O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*, *16*(4), 91–114. https://doi.org/10.1257/089533002320950993

Kansas State Department of Education. (2018). *Kansas ESSA State Plan*. https://www.ksde.org/Agency/Division-of-Learning-Services/Special-Education-and-Title-Services/Every-Student-Succeeds-Act-ESSA. Accessed February 26, 2022.

McEachin, A., & Atteberry, A. (2017). The Impact of Summer Learning Loss on Measures of School Performance. *Education Finance and Policy*, *12*(4), 468–491. https://doi.org/10.1162/edfp_a_00213

McEachin, A., & Polikoff, M. S. (2012). We Are the 5%: Which Schools Would Be Held Accountable Under a Proposed Revision of the Elementary and Secondary Education Act? *Educational Researcher*, *41*(7), 243–251. https://doi.org/10.3102/0013189X12453494

Mead, S. (2016). Don't Forget the Early Elementary Grades. *US News and World Report*. https://www.usnews.com/opinion/articles/2016-10-06/early-elementary-education-years-are-important-for-public-policy. Accessed March 8, 2022.

Mississippi Department of Education. (2019). *Mississippi Consolidated State ESSA Plan*. https://www.mdek12.org/sites/default/files/Offices/MDE/SSE/mississippi-essa-consolidated-state-plan-usde-v6-2019.09-submitted-clean.pdf. Accessed February 26, 2022.

Petrilli, M. J. (2022, January 5). The Case for Kindergarten Tests. *Education Next*. https://www.educationnext.org/case-for-kindergarten-tests-starting-naep-4th-grade-much-too-late/. Accessed March 8, 2022.

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2015). Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps. In *Handbook of Research in Education Finance and Policy* (2nd ed., pp. 491–509). Routledge.

Regenstein, E., Connors, M., & Romero-Jurado, R. (2017a). School Improvement Starts Before School: Under ESSA, States Can Start Re-Orienting Districts Towards the Early Years. *CEELO*. http://ceelo.org/essa_blog_6_reorienting_districts_to_early_years/. Accessed July 10, 2020.

Regenstein, E., Connors, M., Romero-Jurado, R., & Weiner, J. (2017b). *Uses and Misuses of Kindergarten Readiness Assessments* (p. 48). The Ounce. https://startearly.org/app/uploads/pdf/PolicyConversationKRA2017.pdf. Accessed June 9, 2019.

Saw, G., Schneider, B., Frank, K., Chen, I.-C., Keesler, V., & Martineau, J. (2017). The Impact of Being Labeled as a Persistently Lowest Achieving School: Regression Discontinuity Evidence on Consequential School Labeling. *American Journal of Education*, *123*(4), 585–613. https://doi.org/10.1086/692665

Shepard, L. A. (1994). The Challenges of Assessing Young Children Appropriately. *The Phi Delta Kappan*, *76*(3), 206–212.

Soland, J. (2019). Are Schools Deemed Effective Based on Overall Student Growth Also Closing Achievement Gaps? Examining the Black-White Gap in Schools. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://doi.org/10.26300/z9hk-q022.

Solley, B. A. (2007). On Standardized Testing: An ACEI Position Paper. *Childhood Education*, *84*(1), 31–37. https://doi.org/10.1080/00094056.2007.10522967

Terziev, J., & Walsh, E. (2018). *Measuring Progress in the Classroom: How Do Different Student Growth Measures Compare? (Fact Sheet)*. Mathematica Policy Research. https://www.mathematica.org/publications/measuring-progress-in-the-classroom-how-do-different-student-growth-measures-compare-fact-sheet. Accessed February 22, 2022.

Thum, Y. M., & Kuhfeld, M. (2020). *NWEA 2020 MAP Growth: Achievement Status and Growth Norms for Students and Schools*. NWEA. https://teach.mapnwea.org/impl/normsResearchStudy.pdf. Accessed September 6, 2021.

US Department of Education. (2019). *Accountability, State Plans, and Data Reporting: Summary of Final Regulations*. https://www2.ed.gov/policy/elsec/leg/essa/essafactsheet170103.pdf. Accessed January 13, 2020.

Walsh, E., & Isenberg, E. (2015). How Does Value Added Compare to Student Growth Percentiles? *Statistics and Public Policy*, *2*(1), 1–13. https://doi.org/10.1080/2330443X.2015.1034390

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and Student Proficiency in America's Largest School District. *Educational Evaluation and Policy Analysis*, *34*(3), 313–327. https://doi.org/10.3102/0162373712440039

Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review*, *29*(1), 138–146. https://doi.org/10.1016/j.econedurev.2009.07.004

Appendix A: Strengths and Weaknesses of Student Growth Percentiles

In addition to being the most commonly used growth measure across state accountability plans, SGPs offer several theoretical advantages over alternative growth measures like so-called "value-added" models (VAMs) used in much of the econometric literature addressing school performance. SGPs provide an intuitive measure of how much the median or average student in a school grew relative to their academic peers that is easier for parents and policymakers to interpret than VAMs (Betebenner, 2011). Further, unlike VAMs, SGPs do not control for student demographic characteristics which could appeal to policymakers both because it represents a more straightforward approach than VAM estimation (Terziev & Walsh, 2018) and because not including student characteristics may avoid setting lower expectations for different subgroups of students (Walsh & Isenberg, 2015).

However, SGPs are not without their shortcomings. While SGPs may be intuitive, the statistical procedure used to generate SGPs is arguably more complex than the econometric models used to estimate VAMs (Walsh & Isenberg, 2015). Additionally, because they do not control for factors that are outside of schools' control (e.g., families' socio-economic status), authors have expressed concerns that SGPs might yield less accurate indications of school (or teacher) contributions to student growth (Terziev & Walsh, 2018). In the case of schools, not controlling for background characteristics might imply that schools' serving historically disadvantaged student populations will have systematically lower SGPs than those serving more advantaged populations (Walsh & Isenberg, 2015).

Appendix B: Results for Reading and Math Separately

Math Only Results

Figure B1: Schools' Average MAP Growth Scores Across Grade Bands
Math Only

Figure B2: Median SGPs Across Grade Bands
Math Only

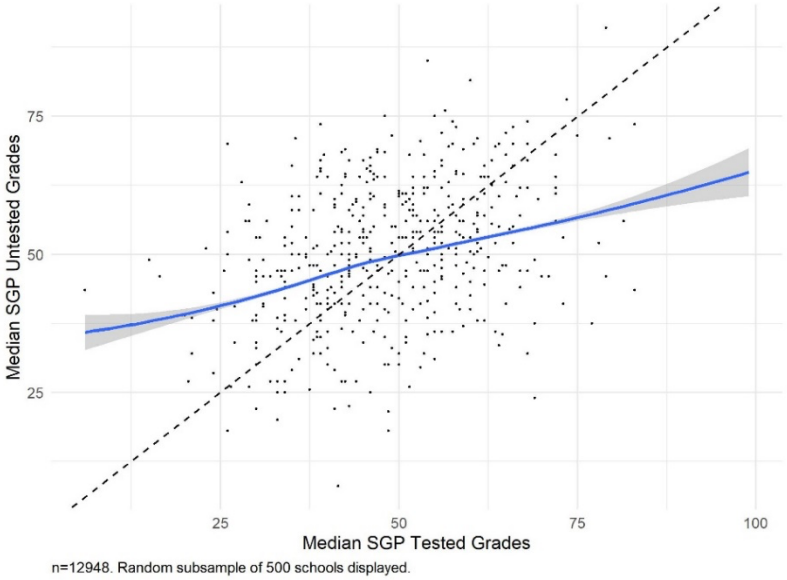


Table B1: Associations Between School Characteristics and Achievement Scores in the Untested Grades - Math Only

	(1)		(2)		(3)	
Achievement Score Tested Grades	0.831	***	0.754	***	0.753	***
% FRPL			-0.002	***	-0.002	***
% Asian					0.003	
% Black					0.002	
% Hispanic					0.002	
% Native American					0.003	
% Pacific Islander					-0.001	
% Two or More Race					0.004	*
% White					0.002	
(Intercept)	0.010	***	0.121	***	-0.107	

n=18,227 school-by-year observations
*p<.10 **p<.05 ***p<.01

Table B2: Associations Between School Characteristics and Growth Scores in the Untested Grades - Math Only

	(1)		(2)		(3)	
Growth Score Tested Grades	0.319	***	0.258	***	0.235	***
% FRPL			-0.105	***	-0.074	***
% Asian					0.243	**
% Black					0.101	
% Hispanic					0.172	
% Native American					0.129	
% Pacific Islander					0.041	
% Two or More Race					0.099	
% White					0.158	
(Intercept)	33.352	***	42.681	***	27.137	**

n=12,948 school-by-year observations
*p<.10 **p<.05 ***p<.01

Table B3: Transition Matrix of School Rating Quintiles Combined Achievement and Growth - Math Scores Only

Quintile (Grade 3-5)	**Quintile (Grade K-5)**				
	1	**2**	**3**	**4**	**5**
1	15.05%	4.27%	0.73%	0.08%	0.01%
2	4.11%	9.61%	5.17%	1.10%	0.06%
3	0.89%	4.92%	8.44%	4.91%	0.85%
4	0.08%	1.20%	4.90%	9.02%	4.73%
5	0.00%	0.05%	0.76%	4.83%	14.21%

n=12,948 school-by-year observations

Table B4: Bottom 5% Matrix for Combined Achievement and Growth - Math Scores Only

	Grades K-5	
Grades 3-5	**Not Bottom 5%**	**Bottom 5%**
Not Bottom 5 %	92.71%	2.04%
Bottom 5%	2.09%	3.16%

n=12,948 school-by-year observations

Reading Only Results

Figure B3: Schools' Average MAP Growth Scores Across Grade Bands - Reading Only

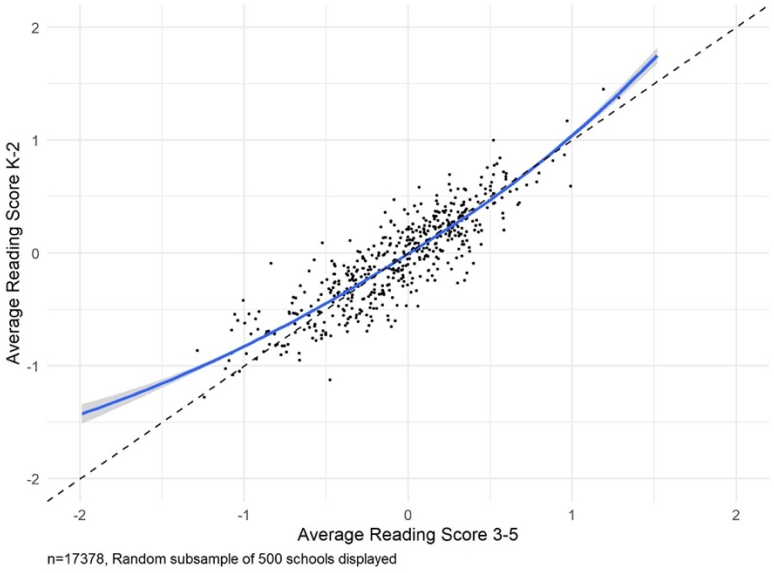


Figure B4: Median SGPs Across Grade Bands
Reading Only

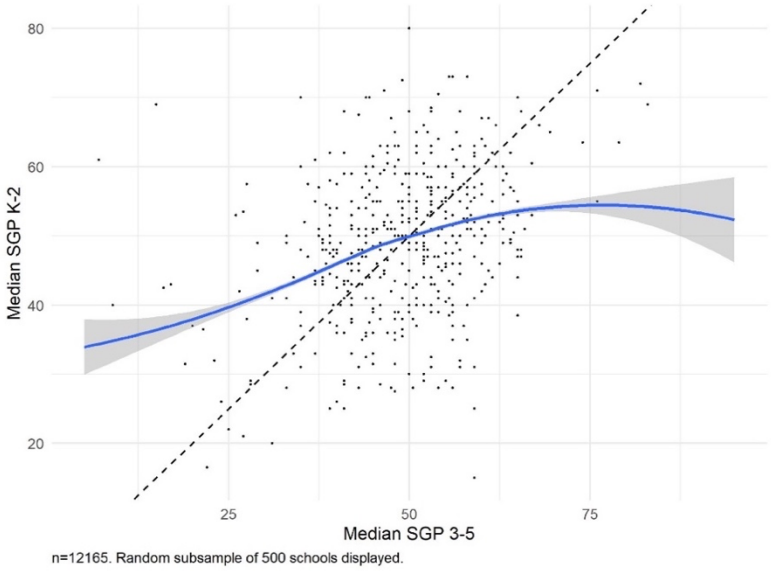


Table B5: Associations Between School Characteristics and Achievement Scores in the Untested Grades - Reading Only

	(1)		(2)		(3)	
Achievement Score Tested Grades	0.898	***	0.829	***	0.827	***
% FRPL			-0.001	***	-0.001	***
% Asian					0.003	
% Black					0.002	
% Hispanic					0.001	
% Native American					0.003	
% Pacific Islander					-0.001	
% Two or More Race					0.003	
% White					0.002	
(Intercept)	0.008	***	0.094	***	-0.074	

n=17,378 school-by-year observations
*p<.10 **p<.05 ***p<.01

Table B6: Associations Between School Characteristics and Growth Scores in the Untested Grades - Reading Only

	(1)		(2)		(3)	
Growth Score Tested Grades	0.338	***	0.240	***	0.218	***
% FRPL			-0.120	***	-0.084	***
% Asian					0.204	**
% Black					0.081	
% Hispanic					0.142	
% Native American					0.116	
% Pacific Islander					-0.125	
% Two or More Race					0.133	
% White					0.142	
(Intercept)	32.501	***	44.496	***	30.379	***

n=12,165 school-by-year observations
*p<.10 **p<.05 ***p<.01

Table B7: Transition Matrix of School Rating Quintiles Combined Achievement and Growth, Reading Scores Only

Quintile (Grade 3-5)	Quintile (Grade K-5)				
	1	**2**	**3**	**4**	**5**
1	14.72%	4.63%	0.74%	0.07%	0.00%
2	4.23%	9.17%	5.27%	1.32%	0.12%
3	1.05%	4.92%	8.12%	5.01%	0.91%
4	0.16%	1.25%	4.95%	8.90%	4.64%
5	0.00%	0.13%	0.94%	4.60%	14.15%

n=12,165 school-by-year observations

Table B8: Bottom 5% Matrix for Combined Scores Reading Scores Only

	Grades K-5	
Grades 3-5	**Not Bottom 5%**	**Bottom 5%**
Not Bottom 5 %	92.91%	1.84%
Bottom 5%	1.86%	3.39%

n=12,165 school-by-year observations

Appendix C: Scatterplots of Achievement and Growth Scores Across Grade Bands

Figure C1: Schools' Average MAP Growth Scores Across Grade Bands Math and Reading

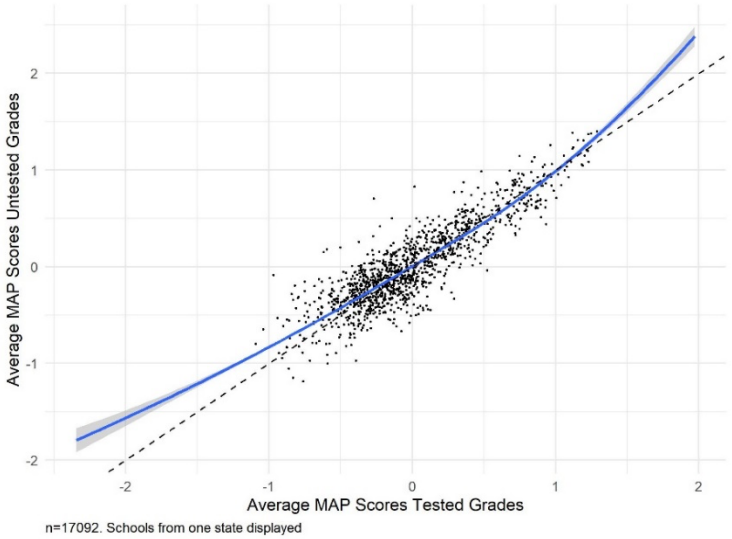
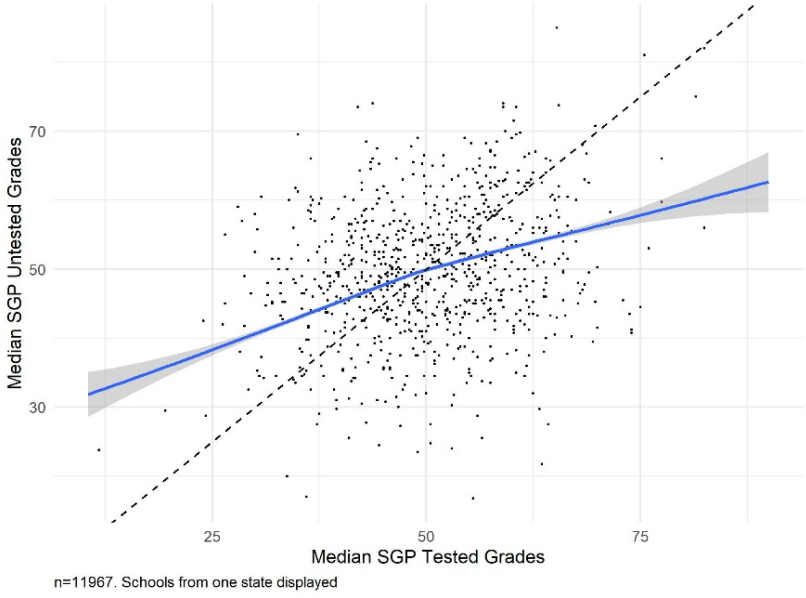


Figure C2: Median SGPs Across Grade Bands Math and Reading



Appendix D: Alternative Weights for Combined Ratings

In the main body of the paper, I report how schools' rankings relative to other schools change when incorporating children's scores from the early elementary grades under a system which weights achievement and growth equally. In this appendix, I report results using alternative weights of achievement and growth.

70% Achievement, 30% Growth

In Table D1, I report the transition matrix indicating the proportion of schools that move quintiles under a school rating system that places greater weight on schools' achievement ratings (70%) relative to their growth ratings (30%). As shown in Figure C1, achievement ratings are more stable across grade bands, and as such fewer schools change quintiles in Table D1 compared to Table 2 in the main body of the text: only 32.8% change quintiles with only 1.5% changing more than one quintile.

Table D1: Transition Matrix of School Rating Quintiles Weighting 70% Achievement, 30% Growth

Quintile (Grade 3-5)	Quintile (Grade K-5)				
	1	**2**	**3**	**4**	**5**
1	16.29%	3.64%	0.22%	0.01%	0.00%
2	3.65%	11.84%	4.26%	0.34%	0.02%
3	0.22%	4.32%	11.21%	4.06%	0.22%
4	0.00%	0.31%	4.16%	11.84%	3.58%
5	0.00%	0.00%	0.18%	3.64%	15.99%

n=12,170 school-by-year observations

30% Achievement, 70% Growth

In Table D2, I report the transition matrix indicating the proportion of schools that move quintiles under a school rating system that places greater weight on schools' growth scores (70%) relative to their achievement scores (30%). Because growth ratings are less stable across grade bands (Figure D2), this weighting system yields comparatively larger changes in schools'

quintile rankings. Table D2 shows that 50.3% of schools change quintile rankings, with 11.2% moving multiple quintiles under a system that places heavier weight on schools' growth ratings.

Table D2: Transition Matrix of School Rating Quintiles
Weighting 30% Achievement, 70% Growth

Quintile (Grade 3-5)	Quintile (Grade K-5)				
	1	**2**	**3**	**4**	**5**
1	13.90%	4.69%	1.31%	0.23%	0.03%
2	4.29%	8.08%	5.34%	2.09%	0.32%
3	1.50%	4.93%	6.85%	5.29%	1.45%
4	0.46%	2.05%	5.11%	7.59%	4.70%
5	0.02%	0.37%	1.41%	4.70%	13.31%

n=12,170 school-by-year observations

Appendix E: Three-Year Average Transition Matrices

The results in the main body of the text reflect changes in schools' relative ranking based on a single year of achievement and growth scores. In practice, though, some states make high-stakes decisions about schools based on three-year averages of their accountability scores. In this vein, Tables E1 and E2 below display transition matrices based on three-year averages of schools combined achievement and growth scores. 38% of schools move quintiles after incorporating the early grades in achievement and growth scores, with 3% moving multiple quintiles. Of schools that fell in the bottom 5% based on test scores in grades 3 through 5, 44% no longer fall in the bottom 5% after incorporating the early grades.

Table E1: Transition Matrix of School Rating Quintiles
Three-Year Averages of Math and Reading Scores

Quintile (Grade 3-5)	Quintile (Grade K-5)				
	1	**2**	**3**	**4**	**5**
1	15.71%	4.17%	0.34%	0.00%	0.00%
2	4.15%	10.86%	4.78%	0.34%	0.00%
3	0.36%	4.36%	10.25%	4.74%	0.29%
4	0.00%	0.69%	4.15%	10.46%	4.59%
5	0.00%	0.04%	0.48%	4.36%	14.87%

n=4,769 observations. Schools' Includes only schools that have three consecutive years of data to calculate three-year averages of achievement and growth scores. Achievement and growth weighted evenly at 50% each.

Table E2: Bottom 5% Matrix for Combined Scores
Three-Year Averages of Math and Reading Scores

	Grades K-5	
Grades 3-5	**Not Bottom 5%**	**Bottom 5%**
Not Bottom 5 %	92.39%	2.33%
Bottom 5%	2.33%	2.96%

n=4,769 observations. Schools' Includes only schools that have three consecutive years of data to calculate three-year averages of achievement and growth scores. Achievement and growth weighted evenly at 50% each.

Appendix F: Robustness Checks

Large School Robustness Check

As Kane and Staiger (2002) discuss, small schools are more likely to have very high or very low achievement and growth scores in a given year because they are particularly sensitive to small changes in the population of students who are included in the achievement and growth calculations. It could be, then, that any movement I observe between quintiles or across the bottom 5% threshold based on the analysis in the main body of the text is driven by small schools. To address this concern, I construct transition matrices for a subset of schools that administer MAP Growth to more than 30 children in each grade level between kindergarten and fifth grade.

Table F1: Transition Matrix of School Rating Quintiles
Large Schools

Quintile (Grade 3-5)	Quintile (Grade K-5)				
	1	**2**	**3**	**4**	**5**
1	14.91%	4.58%	0.69%	0.01%	0.00%
2	4.26%	9.67%	5.06%	1.06%	0.06%
3	0.94%	4.73%	8.56%	4.97%	0.79%
4	0.07%	1.11%	4.88%	9.51%	4.32%
5	0.00%	0.02%	0.81%	4.34%	14.63%

n=9,739 observations. Schools' Achievement and growth weighted evenly at 50% each. "Large" schools are those that administer MAP Growth to more than 30 children in each grade between kindergarten and fifth grade.

Table F1 displays the transition matrix reporting the percentage of schools moving between quintiles after incorporating the early elementary grades into schools' test-based ratings. The table shows that 43% of schools change quintiles with 5% changing multiple quintiles. Table F2 reports that of schools that fell in the bottom 5% with regards to their test-based scores in grades 3 through 5, about 37% no longer fall in the bottom 5% of schools after incorporating the early elementary grades. Both of these figures are very similar to those reported in the main

body of the text, suggesting that my findings are not explained by volatility in small schools' achievement and growth scores.

Table F2: Bottom 5% Matrix for Combined Scores
Large Schools

	Grades K-5	
Grades 3-5	**Not Bottom 5%**	**Bottom 5%**
Not Bottom 5 %	92.71%	1.99%
Bottom 5%	1.97%	3.33%

n=9,739 observations. Schools' Achievement and growth weighted evenly at 50% each. "Large" schools are those that administer MAP Growth to more than 30 children in each grade between kindergarten and fifth grade.

Statistical Noise Simulation

Setting aside the issue of especially small schools, the changes in schools' rankings relative to other schools in their state reported in Tables 4 and 5 in the main body of this paper could be the product of two different phenomena. First, schools may change quintiles because their achievement and growth scores in the early elementary grades differ considerably from their scores in the upper elementary grades. In other words, schools moving between quintiles or above and below the 5% threshold could reflect "true" differences in schools' test-based ratings across the different grade bands incorporated in those ratings.

The second factor, unrelated to schools' actual achievement and growth scores, that could lead to changes in schools' relative performance ranking has to do with sampling variation. As Kane and Staiger (2002) discuss, because elementary schools tend to serve smaller student populations, their test-based ratings are highly sensitive to the specific students included in achievement and growth calculations. In the context of this paper, this suggests that we could see considerable differences in schools' relative ranking across different grade bands even if their "true" achievement and growth scores are the in the early and upper elementary grades simply

because the school's scores are very sensitive to the addition of new students to the achievement and growth calculations.

I conduct a simulation to estimate the changes in schools' rankings that we would observe based on changes to the population of students included in achievement and growth calculations alone rather than "true" differences in mean MAP Growth scores or median SGPs across grade bands. For each school in my data, I take a random sample of students (with replacement) from grades 3 through 5 that is equal in number to the number of students who took the assessment in the same school and year in grades K through 2. These students, on expectation, should have the same average test scores and median SGPs as the children included in the school's original achievement and growth calculations for grades 3 through 5. I then supplement the original sample of students in grades 3 through 5 in that school with this simulated group of students and recalculate each school's mean MAP Growth scores and median SGPs using this new sample of students. As I do in RQ2 in the main body of the paper, I can then average these achievement and growth scores to create a combined score for each school using the simulated data and assign schools to quintiles based on these combined scores.

I then create a transition matrix comparing the quintile ratings of schools in these simulated data to the quintile ratings of the same schools based on the original sample of students' in grades 3 through 5. Any changes in quintiles across the simulated and original data would be attributed to the "noise" introduced by adding new students to the calculations. The results suggest that, based on sampling noise alone, we would expect around 13% of schools to move quintiles when including a group of new students in the achievement and growth calculations (compared to 42% in the main body of the text), with very few schools changing multiple quintiles (0.03% here compared to 5% in the main body of the text). The simulation also

suggests that of the schools that fall in the bottom 5% of schools with respect to their combined achievement and growth scores in the original sample of students in grades 3 through 5, only 11% would move above the bottom 5% threshold after adding the simulated students to the achievement and growth score calculations (compared to 38% in the main body of the text).