COLLABORATIVE FOR STUDENT GROWTH

WORKING PAPER

Reconciling Long-term Education Policy Goals with Short-term School Accountability Models:

Evidence and Implications from a Longitudinal Study

By Jim Soland, Yeow Meng Thum, and Gregory King

Soland, J., Thum, Y.M., & King G. (2019). *Reconciling long-term education policy goals with short-term school accountability models: Evidence and implications from a longitudinal study*. (The Collaborative for Student Growth at NWEA Working Paper).

nwea RESEARCH

Abstract

Schools are increasingly held accountable for their contributions to students' academic growth in math and reading. Under *The Every Student Succeeds Act*, most states are estimating how much schools improve student achievement over time and using those growth metrics to identify the bottom 5% of schools for remediation. These growth determinations are often based on student test scores from two to three years of data. Yet, many objectives ascribed to schools under federal and state policy involve improving much longer-term student outcomes, including preparing students for college. To date, little research has investigated the implications of this discrepancy for school accountability. We begin to close that gap by examining how much rank orderings of schools change when basing estimates of student growth on short- versus long-term timespans. Our results indicate that estimated school effectiveness is highly sensitive to the timespan, suggesting that short-term accountability policies may be generating unintended consequences relative to long-term goals like preparing students for college.

Keywords: school effectiveness, growth modeling, accountability, program evaluation, college and career readiness

Reconciling Long-term Education Policy Goals with Short-term School Accountability Models: Evidence and Implications from a Longitudinal Study

Federal and state accountability policies increasingly hold teachers and schools accountable for contributing to student academic growth in math and reading. At the school level, under *The Every Student Succeeds Act* (ESSA) of 2015, nearly every state plans to use student growth as an accountability indicator in elementary and middle school, oftentimes weighting growth more than achievement estimates (ESSA Plans, 2017). The law also requires states to identify and intervene in the bottom 5% of schools (Council of Chief State School Officers, 2016; Klein, 2016). Oftentimes, these school-level growth estimates (and, therefore, the accountability determinations based on them) model student gains over the course of two to three years (ESSA Plans, 2017). Thus, accountability policies tend to emphasize student academic growth over relatively short time periods.

By contrast, many policies around instruction, curriculum, standards, and assessment emphasize long-term student growth. In particular, college and career readiness is a primary goal under federal and state policy (Conley, 2010). The Common Core State Standards, which are the foundation for assessment and accountability under ESSA in many states, place a high premium on giving students the skills they need to succeed upon college entry (Rothman, 2012). Under ESSA itself, the same law incenting many states to use short-term growth estimates to hold schools accountable, includes provisions emphasizing college and career readiness. Malin, Bragg, and Hackmann (2017) studied provisions under ESSA and provided evidence that the law enacted policies with potential to increase postsecondary preparation.

Thus, state and federal law—even provisions within the same law—result in schools being held accountable for student growth in the short-term, yet the primary educational objectives for those same schools involve growth and development of students over the long-term. To date, few studies examine the consequences of this discrepancy, including implications for which schools are identified as low-performing. We begin to close that gap in the literature by comparing estimates of school contributions to student growth based on (a) fall and spring test scores from a single school year, (b) three years of testing data, and (c) test scores from second grade through the end of elementary school. To help ensure the comparability of the samples used for each set of estimates, we employ the Compound Polynomial (CP) model, which is designed to jointly estimate within- and between-year growth, including school contributions to that growth (Authors, 2018; Thum & Bhattacharya, 2001; Thum & Carl Hauser, 2015; Thum & Matta, 2016). In so doing, we can produce estimates of (a)-(c) using a single model with a consistent set of students.

Our study therefore allows us to investigate two primary research questions about how much estimates of school contributions to student growth differ when:

1. Using fall-to-spring test scores from a single year versus fall-to-spring scores from second through sixth grade?
2. Using fall-to-spring test scores from second through fourth grade versus second through sixth grade?

By answering these two questions, we attempt to show whether accountability policies might produce different results if using longer-term student trajectories (here, grades two through six) versus short-term trajectories. While growth between 2nd and 6th grade is by no means the same as estimating school contributions to student growth between Kindergarten and 12th grade when students could be headed to college, this timespan nonetheless helps show how much estimates of elementary school effectiveness differ when using data from all years that the school serves the student versus only a sample of those years. If different sets of schools are identified as low-performing, then policymakers may need to ask themselves whether accountability policies based on short-term growth are creating incentives that jibe with their broader goal of college and career readiness.

Background

In this section, we briefly describe ways in which federal policies set college and career readiness as a primary goal of the educational system while simultaneously holding schools accountable for short-term growth. Particular emphasis is provided to ESSA given its centrality in the federal accountability landscape.

Policy Emphasis on Long-term Student Outcomes

College and career readiness is increasingly the emphasis of local, state, and federal education policy (Conley, Drummond, de Gonzalez, Rooseboom, & Stout, 2011). The stated aim of the Common Core State Standards, which are used by many states as the basis for their assessment and accountability plans under ESSA, is to define the knowledge and skills students should achieve in order to graduate from high school ready to succeed in entry-level, credit-bearing academic college courses and in workforce training programs (Common Core State Standards Initiative, 2010). According to a study by Conley et al. (2011), students deemed proficient on the Common Core Standards will likely be ready for a wide range of college courses, and that range will widen as students attain proficiency on additional standards.

ESSA itself also emphasizes long-term student outcomes related to college and career readiness (Klein, 2016; Malin et al., 2017). Malin, Bragg, and Hackmann (2017) studied provisions under ESSA and found a strong emphasis on postsecondary readiness, though there is variability in practice due to the high latitude granted to states in implementing the law. According to Malin, Bragg, and Hackmann,

> we discern within ESSA a prominent focus and shift toward [college readiness] as a policy goal. Importantly, this law—which historically focused solely on K-12 education—in many ways now connects K-12 to the higher education sector, including to community colleges. This shift is historically significant and has been underemphasized in the scholarly literature and the media. (2017, p. 828)

Several states are also enacting additional policies related to fostering college and career readiness, in some cases as part of local legislation to implement ESSA (Klein, 2016; Malin et al., 2017).

College and career readiness is not the only long-term outcome emphasized in education policy, either. For example, there are also myriad policies that hold schools accountable for whether students complete high school with the goal of reducing dropout rates (Reardon, Arshan, Atteberry, & Kurlaender, 2010; Roderick, 1994). Further, an aim of many federal education funding streams is to close achievement gaps between white and racial minority students (Lee & Reeves, 2012; Reardon & Robinson, 2008). Oftentimes, the emphasis of related policies is on closing gaps that are present in Kindergarten as students move through school (Quinn, 2015), i.e. on how schools contribute to long-term changes in relative achievement over time.

Short-term Accountability Policies

These long-term student trajectories are not typically mirrored in the policies that hold schools accountable for student achievement (at least in how they are implemented). As the primary federal law governing accountability policy, ESSA gives states much more flexibility to incorporate student growth in achievement into accountability plans than under prior federal law

(Klein, 2016). States have largely responded to this increased flexibility by incorporating growth into school accountability models. Under ESSA, 47 states plan to use student growth as an accountability indicator in elementary and middle school, and 33 states weight student growth the same or more than achievement estimates (ESSA Plans, 2017).

While ESSA does not preclude focusing on long-term growth for accountability purposes, the majority of states estimate school contributions to student growth using test scores from only two years. In most cases, these estimates are produced using traditional value-added models (VAMs), which regress current test scores on a vector of lagged test scores from one or two years prior. Only a handful of states use multiple years of growth beyond two years. For example, Missouri uses a three-year growth model (Missouri State ESSA Plan, 2017) while Arkansas uses a longitudinal model that incorporates as many years of test scores for each student as are available (Arkansas State ESSA Plan, 2017). However states measure growth and weight it relative to static achievement, ESSA requires that states develop a system to identify and improve low-performing schools (generally those deemed to be in the bottom five percent of all schools in the state).

To date, few studies consider how much rank orderings of schools based on estimates of student growth might change depending on the number of years used in the models. Some studies have examined long-term effects of schools on certain student subgroups like English learners (Thomas & Collier, 2002) or the effects of programs like early childhood education on long-term achievement (Barnett, 1995). Studies have also considered contributions of individual teachers to long-term educational achievement (Chetty, Friedman, & Rockoff, 2011). One reason for the sparseness of this literature is likely that modeling student growth using a range of different timespans often results in different samples of students being used due to attrition, student mobility, and other factors (Bates, 2010).

Methods

In this section, we describe our analytic sample, measures used, and modeling strategy, including details of the CP model.

Analytic Sample

We obtained data from a cohort of students in a Southern state that administers tests in math and reading during the fall and spring each year. Table 1 provides descriptive statistics on the students in our sample, who ranged from roughly 86,000 to 139,000 in number depending on the term. Students began in second grade and finished in sixth. We limited the sample to these grades in order to estimate the contributions of schools to students' growth during all of elementary school (excluding Kindergarten and first grade, which are infrequently tested by states). To that end, we assigned students to their modal elementary school. Though a cohort design is employed, the cohort is not intact: students move in and out of the sample so long as they have at least one valid test score.

Figure 1 shows plots of mean achievement by subject and test administration. The purpose of this figure is to illustrate the saw-tooth pattern of achievement. This pattern typically occurs because students see gains in achievement during the school year followed by declines in the summer, often referred to as summer learning loss (McEachin & Atteberry, 2017). Thus, a model designed to estimate trends in fall and spring test scores over time would likely need to account for those seasonal patterns of gain and decline in order to fit the data well (Thum & Hauser, 2015).

We estimated school contributions to student growth for 570 schools altogether. For modeling purposes, we excluded schools serving fewer than 10 students at a given test administration. While our models can be estimated when enrollment is below 10 students, such schools are often anomalous in terms of their students and curricular model. For instance, some of these schools educate students with disciplinary infractions, and may use the test as a placement screener.

One disadvantage of our data considered in the limitations section is that we do not have student covariates often used in the VAM literature. In particular, while we have each student's race, gender, and achievement scores, we do not have socioeconomic, special education, or English learner status. School-level covariates were more complete because we were able to merge our data with those form the National Center for Education Statistics (NCES). Thus, our models included the same covariates as those used by McEachin and Atteberry (2017), including school proportions of white, black, Hispanic, and free or reduced price lunch students. Our models also controlled for total enrollment and whether the school is urban or rural.

Measures Used

In the state we used, virtually all of the students take MAP Growth, an assessment of math and reading. Scores are reported on the RIT scale, which ranges from approximately 120 to 290 and is a transformation of the logit-based Rasch model estimates of student achievement. The tests are vertically scaled, allowing for certain types of growth models to be estimated. MAP Growth was administered in fall and spring, allowing for estimates of fall-to-spring (within-year) and spring-to-spring (between-year) growth. MAP Growth is also computer-adaptive, which means students should receive content matched to their estimated achievement, helping avoid situations where students receive content that is too difficult or easy for them.

Altogether, these attributes of MAP Growth mean we can estimate student growth on a consistent and comparable scale for all time periods in the study.

Using the CP Model

To help address problems of shifting samples of student test takers over time, we used the CP model to simultaneously estimate fall-to-spring (within-year) and spring-to-spring (between-year) growth over five years using a single model. Thus, we were able to compare fall-to-spring growth from a single year to spring-to-spring growth over the course of a student's entire elementary school career, as well as between-year growth using only two years of data (a common practice under ESSA) versus all five years. We relied on properties of conditional multivariate distributions to produce Z-scores that were the basis for our comparisons of school effectiveness, a key element in our strategy to avoid shifts in sample size by timespan being used. Below, we describe the CP design matrix, model, and procedures for estimating conditional Z-scores. These methods are also described in greater detail in prior studies (Authors, 2018; Thum & Bhattacharya, 2001; Thum & Carl Hauser, 2015; Thum & Matta, 2016).

The CP Design Matrix. The biggest difference between the CP and a more traditional, between-year polynomial growth model is the design matrix used. The CP design matrix includes within- and between-year design matrices. To model spring-to-spring between-year growth, the within-year design matrix, $\boldsymbol{D_w}$, is equal to

$$\begin{pmatrix} 1 & -d \\ 1 & 0 \end{pmatrix}$$

Where d is an instructional time interval (e.g., ¾ of a calendar year) that elapses between fall and spring. The between-year design matrix (spring to spring growth), $\boldsymbol{D_b}$, is the same as that of a traditional growth model where $\boldsymbol{D_b}$ =

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix}$$

In $\boldsymbol{D_b}$, there are five rows, one for each year of data, and three columns for the intercept, linear growth term, and polynomial growth term.

Next, we define a 5 x 5 identity matrix, $\boldsymbol{G}$, and calculate the Kronecker product of that matrix with $\boldsymbol{D_w}$ to produce our first CP matrix, **CP1**. That is

$$\mathbf{CP1} = \boldsymbol{G} \otimes \boldsymbol{D_w} . \quad (1)$$

This new matrix, **CP1**, is a 10 x 10 matrix that is equivalent to a piecewise, within-year design matrix with each 2 x 2 diagonal block accounting for a year in the data. We then produce our second CP matrix, **CP2**, using the following Kronecker product with our between-year design matrix, $\boldsymbol{D_b}$**:**

$$\mathbf{CP2} = [\boldsymbol{D_b} \otimes (1{,}0)\,]\,[\boldsymbol{D_b} \otimes (0{,}1)\,] . \quad (2)$$

This function produces a 10 x 6 matrix, **CP2**.

Last, the final **CP** design matrix is produced by multiplying **CP1** and **CP2**:

CP = **CP1** * **CP2** =

$$\begin{pmatrix} 1 & 0 & 0 & -d & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & -d & -d & -d \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 4 & -d & -2d & -4d \\ 1 & 2 & 4 & 0 & 0 & 0 \\ 1 & 3 & 9 & -d & -3d & -9d \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 1 & 4 & 16 & -d & -4d & -16d \\ 1 & 4 & 16 & 0 & 0 & 0 \end{pmatrix} \quad (3)$$

In this matrix, the first three columns represent the intercept, linear growth, and quadratic growth terms for the spring-to-spring portion of the model. Similarly, columns four through six represent the intercept, linear, and quadratic growth terms for the fall-to-spring model. The final CP design matrix used in our study can be found in Table 2.

One should also note that the CP design matrix can be adjusted to center growth at a different grade. For example, as a point of comparison when examining within-year growth estimates to those using all test scores in the data, we centered time in the design matrix at fourth grade (the design matrix presented in Table 2 centers time at second grade). This re-centering is accomplished by subtracting two from the linear spring-to-spring growth column in Table 2 and adjusting the subsequent columns accordingly.

Using the CP Model to Estimate School Effectiveness. The CP model expands traditional growth models to include within-year (fall-to-spring) growth components using the design matrix just described. Given our sample, the CP model allows us to jointly fit between-year growth curve models spanning grades two through six, as well as fall-to-spring gains for each of those years. In our models, X_{tkij} is the *k*th CP growth term for time *t* within student *i* and school *j*:

$$\mathrm{y}_{tij} = \sum_{k=0}^{5} \pi_{kij}\mathrm{X}_{tkij} + e_{tij}. \quad (4)$$

The level-2 model for student *i* within school *j* then becomes

$$\begin{aligned} \pi_{0ij} &= \beta_{00j} + r_{0ij} \quad (5) \\ \pi_{1ij} &= \beta_{10j} + r_{1ij} \\ \pi_{2ij} &= \beta_{20j} + r_{2ij} \\ \pi_{3ij} &= \beta_{30j} + r_{3ij} \\ \pi_{4ij} &= \beta_{40j} + r_{4ij} \\ \pi_{5ij} &= \beta_{50j} + r_{5ij} \end{aligned}$$

Finally, the level-3 model for school j is

$$\begin{gathered}\beta_{00j} = \gamma_{000} + u_{00j} \qquad (6)\\ \beta_{10j} = \gamma_{100} + u_{10j}\\ \beta_{20j} = \gamma_{200} + u_{20j}\\ \beta_{30j} = \gamma_{300} + u_{30j}\\ \beta_{40j} = \gamma_{400} + u_{40j}\\ \beta_{50j} = \gamma_{500}\end{gathered}$$

In the CP model, the first three parameters are comparable to those from traditional growth models. γ_{000} is the predicted spring score in second grade, γ_{100} is the mean school-level linear growth for spring scores, and γ_{200} is the quadratic growth in spring scores across grade levels. The other terms, meanwhile, capture within-year growth. γ_{300} is the predicted fall-to-spring growth in second grade, γ_{400} is how much fall-to-spring growth changes linearly across years, and γ_{500} is the quadratic term for that growth. Thus, the model tells us not only how much within-year growth occurs in the centering grade, but also how we might expect that slope to change as students move through school.

We fit the model in multiple ways treating different coefficients as fixed and random. After testing model fit, (Bentler, 1990; Fieuws & Verbeke, 2006) our preferred model treats all coefficients as random at both the student and school level except for γ_{500}, which is fixed at the school level. Thus, at the school level, or model consists of six parameters, five random and one fixed.

Comparing Estimates of School Effectiveness based on Different Timespans. Using a post-estimation strategy, we were able to estimate school contributions to student growth over different time periods but using the same CP model. We accomplished this objective by employing contrast matrices. For example, to explore the relationships amongst within- and between-year gains in our study, we could use the following contrast matrix $\boldsymbol{C}$

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix}.$$

Using a $\boldsymbol{Y}_i$ matrix of observed RIT scores from spring of second grade, fall of third grade, and spring of sixth grade for student i, the contrast matrix yields:

$$\boldsymbol{C} * \boldsymbol{Y}_i = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} RIT_{fall2nd} \\ RIT_{spring2nd} \\ RIT_{spring6th} \end{pmatrix} = \begin{pmatrix} RIT_{fall2nd} \\ RIT_{spring2nd} - RIT_{fall2nd} \\ RIT_{spring6th} - RIT_{fall2nd} \end{pmatrix}. \quad (7)$$

A similar contrast matrix can also be multiplied with the desired rows from the design matrix **CP.** For example, we can create a new matrix, $\boldsymbol{CP}_{sub}$, that limits **CP** to rows one, two, and ten from Table 2 corresponding to the design matrix for fall of second, spring of second, and spring of sixth grade. We can pair this new subset design matrix with fixed effect estimates from Equation 6 ($\gamma_{000}- \gamma_{500}$) to produce model-based estimates of the above quantities:

$$\boldsymbol{C} * \boldsymbol{CP}_{sub} * \gamma =$$

$$\begin{pmatrix} \widehat{RIT}_{fall2nd} \\ \widehat{RIT}_{spring2nd} - \widehat{RIT}_{fall2nd} \\ \widehat{RIT}_{spring6th} - \widehat{RIT}_{fall2nd} \end{pmatrix}. \quad (8)$$

From there, one can use conventional results for expectations of random variables to estimate the school-level variance-covariance matrix of mean achievement in second grade and of the two growth estimates. Those variances can, in turn, be converted to standard deviations. Thus, for all three rows in the above matrix (Equation 8), one can standardize the observed achievement or growth for a given school (Z score) by subtracting off the model-based estimate of that achievement or growth and dividing by the relevant standard deviation. That is, for a given school:

$$Z = \frac{(RIT_{spring6th} - RIT_{fall2nd}) - (\widehat{RIT}_{spring6th} - \widehat{RIT}_{fall2nd})}{\widehat{SD}_{RIT_{spring6th} - RIT_{fall2nd}}}. \quad (9)$$

This approach is directly comparable to how many achievement and growth norms are produced (Thum & Hauser, 2015). One can even generate these Z-scores conditional on the school's mean RIT in the fall of second grade (or any other starting point). Correlations of these Z-scores then provide information on how much rank orderings of schools change dependent on the timespan used to estimate student growth, all without using a different sample or model. Appendix A provides more detail on how these Z-scores are estimated.

Beyond reporting correlations among these Z-scores based on differing timespans, we also attempted to determine the practical significance of those correlations in an accountability context like ESSA. To do so, we relied on Koedel and Betts (2010), who showed that correlations between VAM estimates lower than .90 will likely result in changes in rank orderings of teachers or schools that alter determinations of effectiveness, including identifying low-performers. Thus, we used .90 as a cutoff for determining practical significance with lower correlations likely indicating changes in which schools might be held accountable under ESSA.

Results

Before turning to results for the two research questions posed, we will first describe the CP model results more generally. Figure 2 presents model-based estimates of RIT scores for each test administration, including spring-to-spring growth trends across those time points. As the figure shows, the saw-tooth pattern in mean test scores generally matches the one shown in Figure 1. Table 3 provides school-level fixed effects estimates from the CP model in math and reading. To make the model parameters clearer to readers, we will interpret the coefficients from math centered at Grade 2 (column 1). Students end second grade with an estimated spring MAP Growth score of 190.6 RIT. Those scores in math are estimated to grow linearly from spring to spring at a rate of roughly 15 RIT per year, though that growth rate slows as students progress through school (quadratic term of -1.78).

Thus far, the three estimated parameters largely match those from a traditional polynomial growth model based only on spring test scores. By contrast, column one also indicates that students increase their test scores between fall and spring of second grade by an estimated 13.4 RIT. However, that within-year growth from year one slows as students move through school at a rate of roughly 1.4 RIT per year. Thus, in practical terms, students tend to make the largest within-year gains during second grade, after which those gains slow, on average.

Question 1. How Much Do Estimates of School Contributions to Student Growth Differ When Using Fall-to-Spring Test Scores from a Single Year Versus Fall-to-Spring Scores from Second through Sixth Grade?

Figure 3 shows plots of mean achievement scores by test administration and subject based on estimated parameters from the CP model. Brackets have been added to these plots to indicate the time spans being used to estimate school contributions to student growth (Z scores) that are being compared. For example, fall-to-spring growth in grades three and four are being compared to spring-to-spring gains from second to sixth grade (the duration of a student's elementary school career for which there are available test scores). As the figure shows, the correlations between within-year growth and estimated growth over the course of elementary school are very low (rho grade 3 and rho grade 4 in the figure). In math and reading, the correlations between Z scores for short- and long-term growth are highest for third grade with estimates of roughly .095. These results indicate that the gains attributable to schools within a single year are only tenuously associated with growth produced during elementary school.

Question 2. How Much Do Estimates of School Contributions to Student Growth Differ When Using Fall-to-Spring Test Scores from Second through Fourth Grade Versus Second through Sixth Grade?

Figure 4 is the same as in Figure 3, but instead compares estimated student growth between spring of second and fourth grade to growth between second and sixth. One should note that, even though these estimates are considering growth between spring scores, these estimates are still based on CP model parameters, which also account for within-year growth. Here, the correlations are .584 in math and .556 in reading. While these correlations are by no means small, they also likely have consequences for which schools would be deemed low-performing under ESSA. As pointed out by Koedel and Betts (2010), VAM-based estimates of teacher and school effectiveness differ in practically meaningful ways under many accountability policies when correlations between estimates dip below .90. Thus, the correlations from our models suggest that results under ESSA will be very different for schools depending on whether two years or four years of growth are used.

Discussion

A conundrum underlies much of state and federal education policy: whereas the aims of many policies are for schools to prepare students for long-term success like finishing high school and obtaining postsecondary training, those same schools are often held accountable for their contributions to student growth over very discrete time periods. To date, little research considers the implications of this conundrum, especially for holding schools accountable for student growth under statutes like ESSA. One reason for this gap in the literature is that there are not many statistical models available that can be used to (a) simultaneously compare long- and short-term growth, including comparing within- and between-year growth, and (b) make such comparisons without shifts in the sample of students being used in estimation.

We begin to close this gap in the literature by employing the CP model, which is specifically designed to be able to jointly estimate between- and within-year growth, including describing trends in the latter. Therefore, we can examine questions relevant to how much rank orderings of schools change depending on whether they are held accountable for short- versus long-term growth. To that end, we produce a few relevant findings.

First, we show that estimates of school contributions to student growth based on fall-to-spring test score gains from a single year are only correlated with estimates of student growth between spring of second and sixth grade at .10 or below. In practical terms, we find little relation between how much a school contributes to within-year growth for a single schoolyear and growth over the course of elementary school. While fall-to-spring growth estimates are not the most common timespan under school accountability law, they are used in some contexts. For example, New York State bases teacher effectiveness determinations on fall-to-spring growth and the effectiveness of several programs have been evaluated using within-year growth (Jensen, Rice, & Soland, 2018).

Second, while correlations between estimated school contributions to student growth based on two years of data versus all five are much higher (generally around .50), they are still low enough that accountability determinations under ESSA would likely look different dependent on which timespan was used. As Koedel and Betts (2010) suggest, determinations made under accountability policies like those under ESSA designed to identify extremely low- and high-performing teachers or schools are likely to differ when estimates correlate below .90. Thus, our correlations of .50 are low enough that different schools would likely be identified as low-performing under ESSA when using short-term growth versus growth during the entire span of elementary school.

Together, our findings suggest that a broader conversation among policymakers about how to hold schools accountable may be warranted. Under federal policy, helping students finish high school, preparing them for college, and closing achievement gaps between white and racial minority students are primary aims of the educational system (Conley, 2010; Klein, 2016; Lee & Reeves, 2012; Malin et al., 2017). All of those goals involve contributions of schools to the long-term growth of their students. While we could not estimate school contributions to student growth between Kindergarten and 12th grade, we were able to estimate contributions to growth over all tested grades in elementary school. The associations between those estimates and the ones based on shorter-term growth—i.e., the durations typically used under ESSA—differed in statistical and practical significance.

On one hand, holding schools accountable for growth rather than statistic achievement (a frequent occurrence under ESSA) is likely to support efforts around college readiness (Reardon, 2016). As Reardon (2016) shows, rank orderings of districts are different when based on achievement versus growth. Further, research suggests that growth is often a better predictor of college readiness than static achievement (Thum & Matta, 2016). On the other, our results provide evidence that the discrepancy between the long-term goals of policy and the short-term nature of accountability may be producing unintended consequences that undermine efforts to prepare students for their futures beyond school walls.

Limitations

This study has several limitations that bear mention. First, our sample is from one state, therefore results may not generalize to the United States. While nearly every student in the state took MAP Growth, there may also be slight differences between our sample and the state's population. Thus, results may not generalize to other states or to the nation as a whole.

Second, MAP Growth is a low-stakes assessment. In the state we used, educators use MAP Growth to monitor student progress and set goals for growth, which can be meaningful for students, yet are not the same as using the score to hold students, teachers, or schools accountable. Therefore, one cannot be sure the same findings would apply to high-stakes contexts. Despite the low-stakes nature of MAP Growth, there is reason to believe that disengagement among students is not primarily responsible for results. Kuhfeld and Soland (2018) showed that estimates of school effectiveness using MAP Growth data change little when results use achievement test scores that correct for rates of disengaged responses among examinees, and Jensen, Rice, and Soland (2018) find little evidence that disengagement on MAP Growth biases estimates of teacher effectiveness.

Finally, comparable to McEachin and Atteberry (2017), we do not have a complete set of student covariates used in much of the VAM literature. In particular, we do not have student-level information on socioeconomic status like free- and reduced-price lunch status. Therefore, we cannot be sure if adding those covariates would change our findings.

Conclusion

There is a conundrum underlying federal education policy: while a primary objective of schools under the law is to prepare students in the long-term for college and career, school accountability determinations are based on short-term growth in achievement. Our results indicate that the schools held accountable using short timespans like those often used under ESSA would likely be quite different if longer timespans were used to estimate school effectiveness. We show this using the CP model, which can jointly model between- and within-year growth, as well as be used to compare estimates of school contributions to that growth using very different timespans. That is, we largely remove concerns that results are due to different sets of students tested across time periods because we base all estimates on a single model with a consistent sample. Our results likely have implications for policymakers, who may wish to engage in a conversation on how best to hold schools accountable for the long-term goals established under law, as well as whether the current emphasis on short-term growth is producing unintended consequences relative to those long-term goals.

References

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children*, 25–50.

Bates, J. K. R. D. (2010). Cross-classified models in the context of value-added modeling.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. National Bureau of Economic Research.

Conley, D. T. (2010). *College and career ready: Helping all students succeed beyond high school*. Hoboken, NJ: John Wiley & Sons.

Conley, D. T., Drummond, K. V., de Gonzalez, A., Rooseboom, J., & Stout, O. (2011). Reaching the goal: The applicability and importance of the Common Core State Standards to college and career readiness. Eugene, OR: Educational Policy Improvement Center.

Council of Chief State School Officers. (2016). *Major provisions of Every Student Succeeds Act (ESSA) related to the education of English learners*. Washington, D.C.

Fieuws, S., & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, *62*(2), 424–431.

Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *20*(1), 90–98.

Klein, A. (2016). The every student succeeds act: An ESSA overview. *Education Week*, 114–95.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education*, *5*(1), 54–81.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990–2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, *34*(2), 209–231.

Malin, J. R., Bragg, D. D., & Hackmann, D. G. (2017). College and career readiness and the Every Student Succeeds Act. *Educational Administration Quarterly*, *53*(5), 809–838.

McEachin, A., & Atteberry, A. (2017). The impact of summer learning loss on measures of school performance. *Education Finance and Policy*, *12*(4), 468–491.

Quinn, D. M. (2015). Kindergarten black–white test score gaps: Reexamining the roles of socioeconomic status and school quality with new data. *Sociology of Education*, *88*(2), 120-139.

Reardon, S. (2016). *School district socioeconomic status, race, and academic achievement*. Palo Alto, CA: Center for Education Policy Analysis at Stanford University.

Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis*, *32*(4), 498–520.

Reardon, S. F., & Robinson, J. P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. *Handbook of Research in Education Finance and Policy*, 497–516.

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal*, *31*(4), 729–759.

Rothman, R. (2012). A common core of readiness. *Educational Leadership*, *69*(7), 11–15.

Thomas, W. P., & Collier, V. P. (2002). A national study of school effectiveness for language minority students' long-term academic achievement. Santa Cruz, CA: University of California at Santa Cruz, Center for Research on Education, Diversity, and Excellence.

Thum, Y. M., & Bhattacharya, S. K. (2001). Detecting a change in school performance: A Bayesian analysis for a multilevel join point problem. *Journal of Educational and Behavioral Statistics*, *26*(4), 443–468.

Thum, Y. M., & Carl Hauser. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth.* Portland, OR: NWEA.

Thum, Y. M., & Matta, T. (2016). Fitting curves to data series with seasonality using the Additive Polynomial (AP) growth model. Presented at the the National Council on Measurement in Education, Washington, D.C.

Table 1

Analytic Sample Descriptive Statistics

Year	Term	Grade	Test Admin.	Student Count	Race & Gender Proportions				Mean Achievement (RIT Scale)	
					Black	Hisp.	White	Female	Math	Reading
2010	Fall	2	1	85,674	0.332	0.076	0.509	0.511	178.170	175.333
2011	Spring	2	2	92,122	0.331	0.075	0.512	0.511	191.990	189.367
2011	Fall	3	3	103,625	0.331	0.074	0.513	0.510	192.503	190.728
2012	Spring	3	4	104,614	0.331	0.073	0.514	0.510	204.730	200.479
2012	Fall	4	5	131,569	0.329	0.074	0.514	0.510	203.925	200.299
2013	Spring	4	6	132,748	0.328	0.075	0.515	0.510	213.651	207.436
2013	Fall	5	7	112,068	0.325	0.075	0.517	0.510	211.804	206.563
2014	Spring	5	8	138,896	0.326	0.076	0.516	0.508	220.796	212.596
2014	Fall	6	9	123,698	0.333	0.077	0.524	0.511	216.514	211.448
2015	Spring	6	10	125,545	0.330	0.079	0.525	0.511	222.030	215.060

Table 2

Spring to Spring Cumulative Polynomial Design Matrix Centered at Grade 2

Year	Term	Grade	Test Administration	Design Matrix: Spring to Spring Growth			Design Matrix: Fall to Spring Growth		
				Intercept	Linear	Quadratic	Intercept	Linear	Quadratic
2010	Fall	2	1	1	0	0	-1	0	0
2011	Spring	2	2	1	0	0	0	0	0
2011	Fall	3	3	1	1	1	-1	-1	-1
2012	Spring	3	4	1	1	1	0	0	0
2012	Fall	4	5	1	2	4	-1	-2	-4
2013	Spring	4	6	1	2	4	0	0	0
2013	Fall	5	7	1	3	9	-1	-3	-9
2014	Spring	5	8	1	3	9	0	0	0
2014	Fall	6	9	1	4	16	-1	-4	-16
2015	Spring	6	10	1	4	16	0	0	0

Table 3

Fixed Effects Estimates from CP Growth Models

	Math		Reading	
	Centered at 2nd Grade	Centered at 4th Grade	Centered at 2nd Grade	Centered at 4th Grade
	(1)	(2)	(3)	(4)
1. Intercept - between year	190.561	212.428	188.256	206.224
	0.244	0.288	0.267	0.250
2. Linear - between year	14.497	7.370	11.765	6.203
	0.115	0.044	0.088	0.031
3. Quadratic - between year	-1.782	-1.782	-1.390	-1.390
	0.026	0.026	0.021	0.021
4. Intercept - within year	13.432	10.055	13.745	7.331
	0.123	0.087	0.121	0.061
5. Linear - within year	-1.381	-1.995	-3.953	-2.461
	0.110	0.032	0.104	0.034
6. Quadratic - within year	-0.153	-0.153	0.373	0.373
	0.025	0.025	0.023	0.023

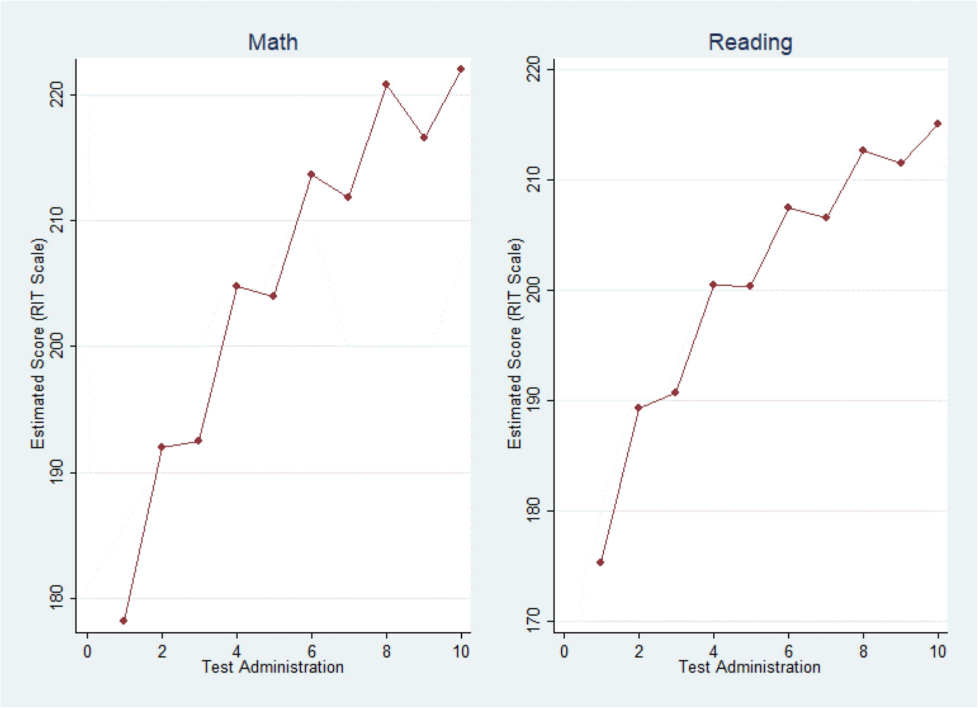

Figure 1. Scatterplots of mean RIT scores by test administration and subject.

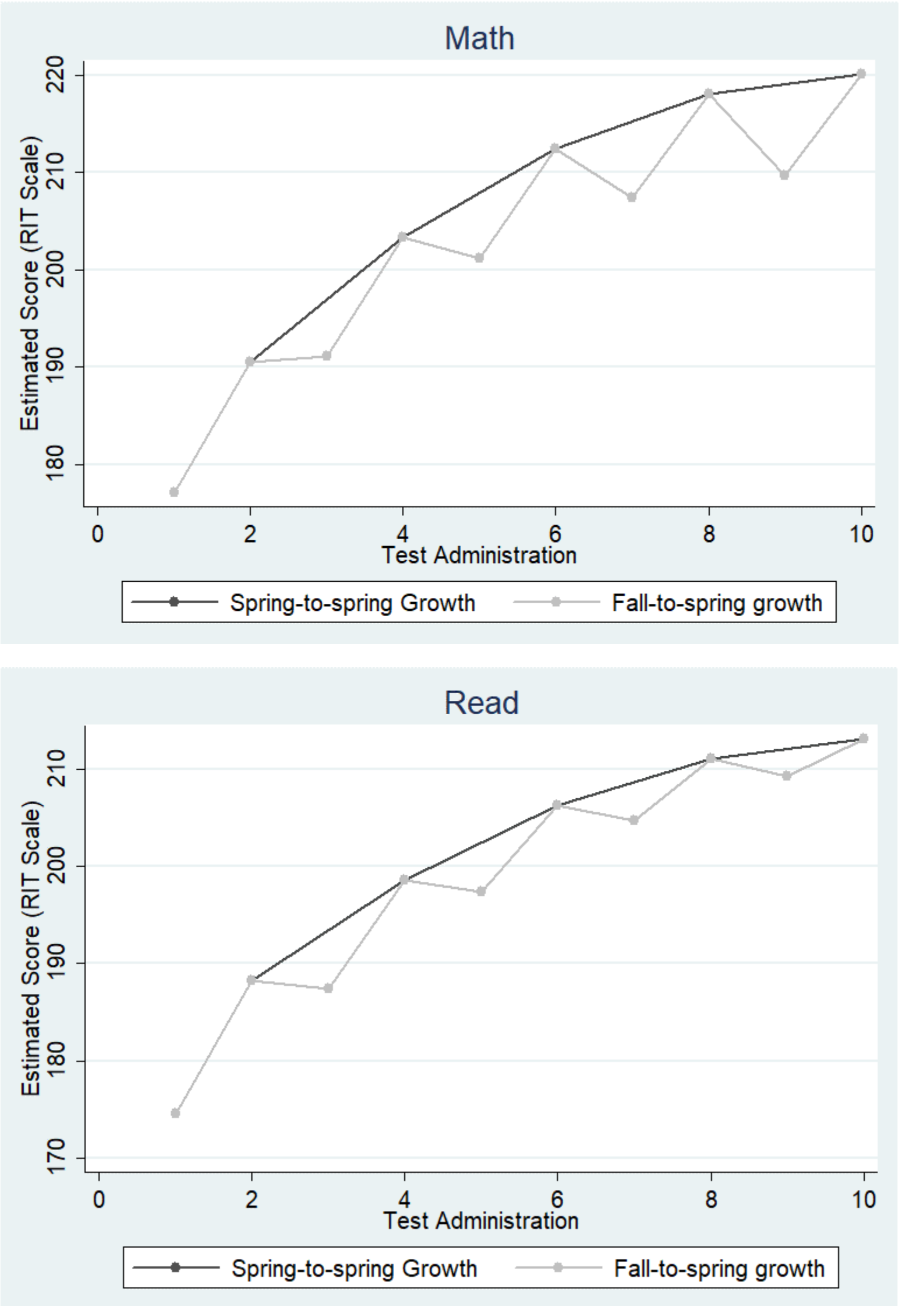

Figure 2. Model-based plots of estimated RIT scores by subject and test administration.

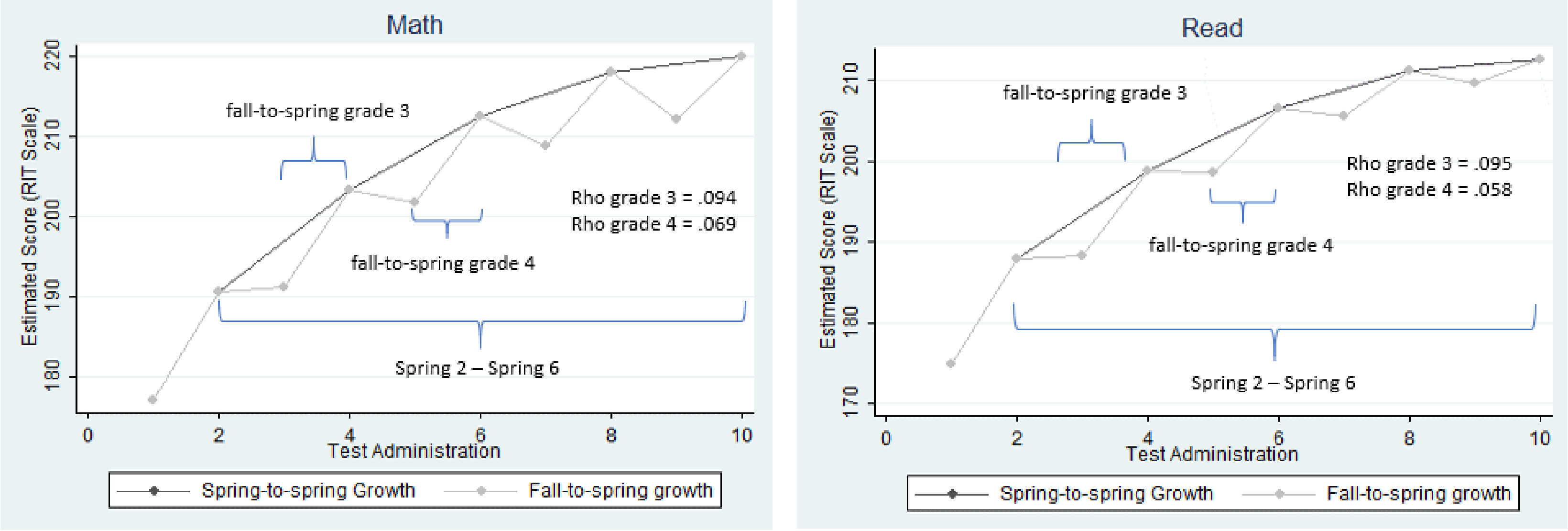

Figure 3. Plots of time intervals used to compare school contributions to within-year growth and school contributions to growth during elementary school.

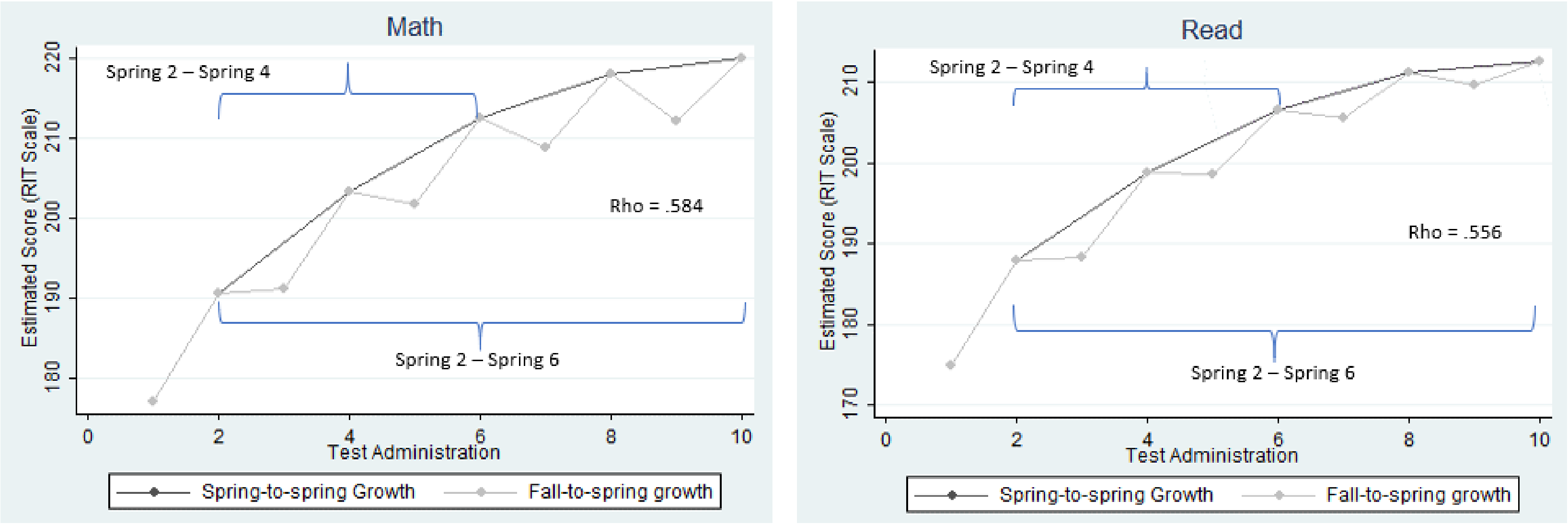

Figure 4. *Plots of time intervals used to compare school contributions to growth between spring of second and fourth grade with contributions to growth between spring of second and sixth grade.*

Appendix A. Generating Z-scores Based on Random Effects Distributions

Suppose that student i receives pre-test and post-test scores $Y_i = [Y_t, Y_{t+1}]$. Following the development of prediction results for the multilevel growth model given by Thum and Hauser (2015), we can define a contrast matrix such that:

$$\boldsymbol{C} = \begin{pmatrix} c_1' \\ c_2' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \quad \text{(A1).}$$

Using this contrast matrix, we can produce

$$\boldsymbol{CY} = \begin{pmatrix} Y_t \\ Y_{t+1} - Y_t \end{pmatrix} = \begin{pmatrix} Y_t \\ G \end{pmatrix} = \boldsymbol{P} \quad \text{(A2)}$$

where Y_t is the starting RIT and G is the gain. The predicted achievement is $\hat{y} = c_1'\boldsymbol{A}\hat{\gamma}$ and the marginal predicted gain between times 1 and 2 then becomes $\hat{G} = c_2'\boldsymbol{A}\hat{\gamma}$, where $\hat{\gamma}$ is the $1\ x\ 6$ vector of school fixed effects estimates from Equation 8 and $\boldsymbol{A}$ is a matrix composed of the rows from the appropriate design matrix corresponding to the time points of interest (see Table 2). For example, when looking at growth between fall and spring of 2nd grade using spring-to-spring between-year growth centered at grade two, $\boldsymbol{A}$ would correspond to the first two rows of the design matrix in Table 2. Conventional results for expectations of random variables give the standard error of $\hat{G}$ as

$$se(\hat{G}) = \sqrt{c_2'\boldsymbol{A}Var(\hat{\gamma})\boldsymbol{A}'c_2} \quad \text{(A3).}$$

We can similarly estimate the school-level variance-covariance matrix of Y_t and gain G as

$$\boldsymbol{V}_s = \boldsymbol{C}\left[\boldsymbol{A}Var(\hat{\gamma})\boldsymbol{A}' + \boldsymbol{AH}_3\widehat{\boldsymbol{T}}_\beta\boldsymbol{H}_3'\boldsymbol{A}'\right]\boldsymbol{C}'. \quad \text{(A4)}$$

Here, $\boldsymbol{H}_3$ is a selection matrix that identifies the random coefficients among $\boldsymbol{\beta}_j$ for schools with estimated variance-covariances of $\widehat{\boldsymbol{T}}_\beta$. As an example, if a model included six fixed effects but only five random effects at the school level (as ours does), $\boldsymbol{H}_3$ selects from $\boldsymbol{T}_{\boldsymbol{\pi}}$ (see Equation 4) the terms corresponding to the five parameters with random effects.

From here, one can take the square root of $\boldsymbol{V}_s$ to get the standard deviation of achievement and gain estimates. $\hat{G}$ can then be subtracted from an observed, school-level average gain between times one and two, and that whole value divided by the standard deviation to produce a Z-score, a growth effect size, for the gain between two time points. That is, we estimate a Z-score for where a given school's mean observed gain falls relative to the model based mean and standard deviation of that gain. Those Z-scores can then be correlated across models to determine how rank orderings of schools might change dependent on the test administrations used to estimate growth.

One can also take the above approach and estimate the same Z-score, but do so conditional on a given school's starting RIT, another empirically-anchored growth effect-size introduced as the "conditional growth index," or CGI, in Thum and Hauser (2015). This conditioning better accounts for the fact that school-level growth may be correlated with initial mean achievement. The method is also more akin to various baseline VAMs that condition on an initial pretest score. For a given school j with starting RIT $\bar{y}_{j1}$, the expected conditional gain can be expressed as

$$G_{Sj}^{*} = \hat{G} + \boldsymbol{V}_{s[2,1]} \cdot \boldsymbol{V}_{s[1,1]}^{-\mathbf{1}} \cdot \left(\bar{y}_{j1} - \hat{y}\,\right) \quad \text{(A5)}$$

And the expected conditional standard deviation for those gains as

$$SD_{sj}^{*} = \sqrt{\boldsymbol{V}_{s[2,2]} - \boldsymbol{V}_{s[2,1]} \cdot \boldsymbol{V}_{s[1,1]}^{-\mathbf{1}} \cdot \boldsymbol{V}_{s[1,2]}} \qquad \text{(A6).}$$

G_{Sj}^{*} and SD_{Sj}^{*} can then be used to produce Z-scores just as before. This is the approach that we used when producing Z-score correlations across estimates of school contributions to student growth.

ABOUT THE COLLABORATIVE FOR STUDENT GROWTH

The Collaborative for Student Growth at NWEA is devoted to transforming education research through advancements in assessment, growth measurement, and the availability of longitudinal data. The work of our researchers spans a range of educational measurement and policy issues including achievement gaps, assessment engagement, social-emotional learning, and innovations in how we measure student learning. Core to our mission is partnering with researchers from universities, think tanks, grant-funding agencies, and other stakeholders to expand the insights drawn from our student growth database—one of the most extensive in the world.