School Effectiveness, Summer Loss,

and Federal Accountability:

Apply the Compound Polynomial Model in a Program Evaluation Context

By Jim Soland and Yeow Meng Thum

nwea RESEARCH

Abstract

Under *The Every Student Succeeds Act* (ESSA) of 2015, school are being held accountable for their contributions to student growth in math and reading achievement. Meanwhile research shows that estimates of school effectiveness are sensitive to whether they account for the time students spend out of school during the summer. Despite the importance of student growth under ESSA and evidence on how summer learning loss can impact estimates of school effectiveness, most statistical models used in research and accountability do not account for the seasonality of achievement data. In this study, we apply the Compound Polynomial or "CP" model in a school evaluation context. The CP model addresses the seasonality of student test scores by simultaneously estimating between- and within-year growth. By presenting the CP in this context, we provide a new statistical model that can be used to estimate school effectiveness in the presence of seasonal data. From a policy standpoint, we produce evidence on how much ignoring summer loss may impact school accountability determinations under ESSA and other accountability frameworks that draw evidence from trends in assessment data.

Keywords: school effectiveness, growth modeling, seasonality, summer loss, program evaluation

School Effectiveness, Summer Loss, and Federal Accountability:

Applying the Compound Polynomial Model in a Program Evaluation Context

Research suggests that estimates of school and district quality based on achievement are very different than when based on estimates of growth in achievement over time (most recently, Reardon, 2017). Perhaps based in part on this body of research, The Every Student Succeeds Act (ESSA) of 2015—the primary law governing federal education accountability—emphasizes holding schools accountable for their contributions to student growth in math and reading over time. Under the new law, 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 static, point-in-time achievement estimates (ESSA Plans, 2017). Further, under the law, policymakers will often use these growth estimates to identify and intervene in the bottom 5% of schools in any given state (Council of Chief State School Officers, 2016; Klein, 2016). Thus, the stakes for estimating student growth in a reliable and justifiable way are high.

Despite major policy emphasis on student academic growth, no accountability plans under ESSA we are aware of account for summer learning loss in their models. Summer loss is a well-documented phenomenon: students tend to produce test score gains, on average, between fall and spring, while achievement tends to drop off during the summer when students are not in school (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017). This issue is typically ignored in federal accountability (and oftentimes in program evaluation as well) despite research showing that this oversight can have practical consequences for which schools are deemed effective or ineffective (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017). Oftentimes, rank orderings of schools based on their contribution to growth shift substantively when using

spring-to-spring versus fall-to-spring growth (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017).

There are two primary reasons summer loss is often overlooked in federal accountability and program evaluation. First and foremost, many states test students only once per year, eliminating the possibility of estimating within-year growth. A second reason is the lack of models designed to understand within-year (e.g., fall-to-spring) versus between-year (e.g., spring-to-spring) growth. On one hand, some innovative models have been employed to examine summer learning loss, including its impact on estimates of school and teacher effectiveness (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017; von Hippel, Workman, & Downey, 2017). On the other, many of these studies address seasonality in student testing data by estimating school contributions to growth separately by year, using lags rather than a true growth model, or employing nonparametric approaches that are agnostic on the functional form of trends in student data over years (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017; von Hippel et al., 2017).

Further, some of these models are limited because the sample of students being tested shifts between fall and spring (Gershenson & Hayes, 2018; McEachin & Atteberry, 2017). To address this issue, researchers often limit estimates of school effectiveness only to students with test scores at both time periods, an approach that is not practicable for accountability purposes. This sample problem is exacerbated even more when fitting growth models that span several years and comparing them to fall-to-spring estimates from a given year.

To help close this gap in the literature and provide evaluators with an additional modeling option, we apply the Cumulative Polynomial (CP) model developed by Thum and Hauser (2015) for estimating school contributions to student growth. Thum (2018) examined the CP as an

instance of the general approach of adding multiple suitably chosen curve components, of which the familiar piece-wise polynomial is an example. In this application, the CP model provides a means to simultaneously fit within- and between-year growth sub-models and, therefore, to better account for seasonality (Thum & Hauser, 2015; Thum & Matta, 2016). We fit this model using a dataset that includes vertically scaled test scores from fall and spring. The combination of our model and dataset means we can investigate the effect of seasonality on estimates of school contributions to student growth in ways unique to this literature. Further, by fitting our growth model rather than a value-added model (VAM) that regresses post-test scores on pre-test scores like those of McEachin and Atteberry (2017) and Gershenson and Hayes (2018), we can estimate patterns of within-year growth over several years that allow for more direct comparisons of school rankings from long-term growth models compared to single-year estimates more typical in the VAM literature.

Using our data and growth model, we investigate three research questions. First, how much do students' within-year gains shift over time as they move through school? Practically, this question helps show how much estimates of within-year school effectiveness may be sensitive to the grade levels schools serve. Second, how much of the variance in growth is within schools for fall-to-spring versus spring-to-spring estimates? Third, how strongly correlated are estimates of school effectiveness that use fall-to-spring versus spring-to-spring estimates? Across all three research questions, our primary motivation is not to settle any of these issues. Rather, we intend to explore potential uses, strengths, and weaknesses of the CP model in the context of school evaluation with the broader aim of supplying a new tool that researchers and evaluators can use to better account for the seasonality of student achievement.

Background on Summer Learning Loss and School Effectiveness

Although the literature investigating the impact of seasonality on estimates of school effectiveness remains sparse, the available studies indicate that school effectiveness estimates are sensitive to whether summer loss is accounted for in the model (McEachin & Attebery, 2017; Gershenson & Hayes, 2018; Papay, 2011; von Hippel et al., 2017). These studies have employed different data and it therefore follows that different value-added models (VAMs) have been used.[1] Generally, when repeated test scores are available for a student, this literature relies on two broad categories of VAMs, lag-score and growth models, both of which we detail below in the context of summer loss and school quality research. We then discuss why, given the data we employ, the CP model is likely useful in the context of evaluating school contributions to student growth in the presence of seasonality.

Summer Loss Research Using a Lag-score Model

Lag-score models that regress current test scores on prior test scores, sometimes from multiple subjects, are often used in value-added analyses. For example, performance during prior time periods is accounted for in these models as lags in the production function (Loeb, Soland, & Fox, 2014). Because the slope estimates of these models are insensitive to linear transformations of the pre-test scale score (Briggs, 2013; Briggs & Domingue, 2012; Briggs & Weeks, 2011), lag-score models are particularly useful when scores across time are not vertically scaled (Soland, 2017). Although the school-specific estimates in such models are regression residuals and thus share the same scale as the original test score, they are often reported in standardized units with a mean of zero and variance of one (Raudenbush, 2004). While the test scores used in a lag-score model need not be vertically equated, an assumption that the test score scaled used is equal-interval still applies to the comparisons of residuals, even when standardized (Soland, 2017). This equal-interval assumption is one reason that some research recommends

using value-added models that assume the test scale is ordinal rather than interval (Betebenner, 2009).

Growth in student test scores over time for lag-score models is also not parameterized as it would be in a growth model. Instead, the primary assumption made about the relationship between pre- and post-tests is that the latter are a function of the former, not that there is a specific functional form to growth over time (Raudenbush, 2004; Reardon & Raudenbush, 2009). Given the lack of a functional form for growth in the lag-score model, if scores across grade levels are indeed vertically scaled, and the scores possess equal interval properties, its use likely means that otherwise useful information in the data could be ignored (Kolen, 2011; Thum, 2015a). Further, such models are statistically problematic when the pre-test scores are measured with error (which they invariably are), leading to a robust errors-in-variables VAM literature (Thum, 2003; Lockwood, McCaffrey, & Savage, 2016).

The most prominent lag-score study of school effectiveness in relation to summer loss was conducted by McEachin and Atteberry (2017) who, like us, used a dataset that included Measures of Academic Progress (MAP) Growth scores in math and reading from fall and spring administrations. They quantified the bias produced by seasonality by comparing estimates of spring-to-spring gains to the alternative results for fall-to-spring gains using a fixed-effects lag-score model. Their results made two broad contributions to the literature. First, they quantified the bias introduced into VAM estimates of school effectiveness by ignoring summer learning loss, with the standard deviation of the bias roughly equivalent to 25% of schools' VAM scores using a spring-to-spring model. Second, they showed that the practical effect of this bias is oftentimes an understatement of estimated effectiveness for schools serving low-income children when they are evaluated by their spring-to-spring gains.

Beyond school effectiveness, studies have also looked at the impact of seasonality on teacher VAM estimates using the lag-score model. For example, Gershenson and Hayes (2018) used The Early Childhood Longitudinal Study (ECLS) to estimate teacher-level VAMs relying on spring-to-spring scores, then compared estimates to the results from fall-to-spring models. As in the school effectiveness literature, they found substantive differences in the estimates, with the largest changes in mathematics. Neither Gershenson and Hayes (2018) nor McEachin and Atteberry (2017) used a multilevel model.

Growth Models

When repeated test scores share a common scale, value-added analyses can also be developed using student level growth models, with test scores nested within the student or, as is frequently the case in educational assessment practice, students also nested within schools or districts (Raudenbush, 2004; Thum, 2003). In a growth modeling VAM framework, all test scores are dependent variables, which means they are treated equitably as outcomes in terms of the information they contribute and the role they play in helping us understand the growth of a student. Thus, some of the errors-in-variables corrections are less relevant, though measurement error still remains an issue (Thum, 2003; Lockwood, McCaffrey, & Savage, 2016). The predicted outcomes are also on the same scale, as are the residuals.

Unlike lag-score models, growth models take full advantage of the equal interval properties of the underlying test scale, if present (Briggs, 2013; Thum, 2015).[2] If one assumes a practical vertical scale has been developed, estimates of school effectiveness could be reported on the original scale (Briggs, 2013; Soland, 2017; Thum, 2015a). When growth models are deployed under such scaling conditions, estimates of student growth (and school contributions to it) can be made between any two time points in the sample, and growth can be based on the

specific amount of time that elapses between test administrations when such calendar data are available (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Thum, 2003).

Two primary studies we are aware of use growth models to estimate summer loss, including its impact on school effectiveness. First, Atteberry and McEachin (2015) found statistically significant variability in students' summer growth rates, regardless of grade level. That article also used MAP Growth data. Second, Downey, Von Hippel, and Hughes (2008) estimated multilevel growth models and found differences in school effectiveness dependent on whether spring-to-spring achievement test scores were used compared to fall-to-spring within-year estimates of growth. Similar to the results produced by McEachin and Atteberry (2017), these differences were especially pronounced for schools educating low-income students (Downey et al., 2008).

In both studies, the models used were nearly or fully saturated in that the number of parameters approached the limit of time points available. For example, Atteberry and McEachin had one degree of freedom (10 time points and 9 parameters). Similarly, Downey et al. (2008) employed four parameters to represent changes among performance over four time points. As a result, there is no model specification error in their model and the residual is assumed to be known from the reported reliability of the test scale. Such an approach to the description of change can be regarded more as a reparameterization of a set of repeated outcomes, as is sometimes the practice in doubly multivariate linear models (Johnson & Wichern, 2007), and less represents an effort to determine the functional form underlying the observed trend in the data series, a common objective of growth modeling.

The CP Growth Model

Thum and Hauser (2015) employed a flexible functional form termed the "Compound Polynomial" (CP) for describing growth trends with marked seasonal patterns. The CP represents a new approach by adding suitably chosen polynomials for fitting seasonally varying growth trends (Thum, 2018). The CP model has also been employed to explicitly account for seasonal patterns in achievement growth data in predicting college and career readiness benchmarks keyed on the Scholastic Aptitude Test (SAT) and ACT from middle school achievement in the presence of self-selection in taking the college entrance tests (Thum, 2015b; Thum & Matta, 2015). However, the CP has yet to be introduced into the school effectiveness or summer loss literature.

Analytically, the CP combines separate depictions of (1) a polynomial growth in performance (scores) that occurs within a segment (generically for year or grade level) and (2) a second polynomial growth model for one or more within-segment growth components (e.g., the predicted fall score or the linear growth rate for the year or grade level) to arrive at a distinctive reparameterization of the growth in test scores over time for the student. That is, the model includes parameters capturing between- and within-year growth. Further, the CP model can be expanded to include as many within-year test administrations as available, and employ instructional times that elapse between tests unique to the student.

Prior evidence suggests that the CP model fits test score data from multiple time points within a year (e.g. fall, winter, and spring) better than traditional polynomial models in the presence of marked seasonality. For example, Thum and Matta (2015) and Thum and Hauser (2015) provided evidence that the CP fit better than a traditional polynomial model that ignored the seasonality of the data, as measured by the Bayesian Information Criterion (BIC) and overall residual variance of the model. Further, results indicated that traditional polynomials tended to

overpredict test scores in the fall and under predict them in the spring, as one might expect given seasonal rising and falling in the data. Errors for analyses using a traditional polynomial exhibit strong auto-correlations as a result. Under such conditions, inferences from an independently and identically distributed (i.i.d.) error model are thus biased.

In sum, most research examining the effect of summer loss on estimates of school effectiveness use either a lag-score or a growth model. While the lag-score model has some benefits including being insensitive to linear transformations of the pre-test scale score, school-level estimates are residuals and therefore do not allow one to test the functional form of between- and within-year growth. Among studies using growth models to estimate school contributions to student growth in the presence of seasonality, those models are often fully saturated, leaving no degrees of freedom for model fitting or testing. Given the data we employ and the psychometric properties of its outcomes, and the goal of describing and explaining the trend in changes in performance over time, the CP appears to be well-suited for exploring school effectiveness in a way that accounts for summer loss, allows one to test hypotheses about the nature of student growth in achievement, and includes parameters capturing spring-to-spring and fall-to-spring growth.

Methods

Analytic Sample

Our analytic sample consists of one cohort of students in an East Coast state that uniformly administers MAP Growth achievement tests in math and reading at multiple points during the year, including fall and spring. Table 1 presents descriptive statistics on the students in our sample. Students begin in second grade and finish in sixth grade. We limited the sample to these grades so that we could estimate the contributions of students' elementary schools to

their growth in achievement between second grade (the first year tests are typically administered in many ESSA plans) and sixth grade, which is often the first year of middle school. For simplicity, we assigned students to their modal elementary school.[3] While we used a cohort design, the cohort is not intact: students can move in and out of the sample at any time so long as they have at least one valid test score.

Figure 1 plots mean RIT scores in math and reading for the test administrations in Table 1. These mean achievement patterns are distinctly seasonal. In virtually every year, gains between fall and spring are followed by decreases between spring and the subsequent fall (though, in some cases, there is a deceleration in growth between spring and fall rather than a decline in achievement).   As shown in Thum (2003), Thum and Hauser (2015), and Thum and Matta (2015), simply fitting a smooth polynomial to these data would likely lead to an underestimate of mean scores in the spring and an underestimate of those in the fall.

We examined school contributions to student growth for 570 schools in our sample. In our analyses, we excluded schools serving fewer than 10 students in a given grade. While our models can be estimated when enrollment is below 10 students, such schools are often anomalous in terms of their focus or student body. For example, several of these schools were for students with disciplinary problems, and likely used the test as a placement screener.

One disadvantage of using our dataset (described more in the limitations section) is that we do not have several student-level variables commonly used in models designed to estimate school contributions to student growth. Specifically, we have each student's race, gender, school attended, and vectors of achievement scores. We do not have student socioeconomic, special education, or English-learner status. School-level covariates were included by merging our data with the those produced by the National Center for Education Statistics (NCES). Thus, we used

the same covariates as McEachin and Atteberry (2017), including proportions of white, Hispanic, black, and low-income student in each school. Our models also controlled for total school enrollment provided by NCES, and whether the school is deemed urban versus 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 roughly 120 to 290 and is a transformation of the logit-based Rasch model estimates of student achievement. The tests are vertically scaled, which means growth models that support numerical comparisons on the scale of the outcome variable can be estimated across time points. MAP Growth is often administered in fall and spring terms, allowing for estimates of within- and between-year growth. Additionally, MAP Growth is a computer-adaptive test, which means students in any given grade and year should primarily be receiving content that is matched to their estimated achievement level, helping avoid instances where students are receiving content that is extremely difficult or easy for them. In tandem, these attributes of MAP Growth mean that we should be able to estimate student growth on a consistent and comparable scale for all time periods and grades in the study.

Models

In this section we describe methods for estimating school contributions to student growth before discussing methods specific to each research question.

Estimating school contributions to student growth. We estimated school contributions to student growth using models that include three levels, with time points nested within students nested within schools. Before turning to the CP, we will describe a standard growth curve model as a point of comparison. Under such an approach, one could model RIT scores such that y_{tij} is

the math or reading test score for student i in school j at time t (year/term). In these baseline growth models, time t corresponds to the test administration in Table 1. The baseline polynomial model is:

$$y_{tij} = \pi_{0ij} + \pi_{1ij}time + \pi_{2ij}time^2 + e_{tij}. \quad (1)$$

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

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \quad (2)$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij}$$
$$\pi_{2ij} = \beta_{20j} + r_{2ij.}$$

Finally, the level-3 model for school j is

$$\beta_{00j} = \gamma_{000} + u_{00j} \quad (3)$$
$$\beta_{10j} = \gamma_{100} + u_{10j}$$
$$\beta_{20j} = \gamma_{200} + u_{20j}.$$

Variance components of the model are as follows:

$$e_{tij} \sim N\left(0, \sigma_{tij}^2\right) \quad (4)$$

$$\boldsymbol{r}_{ij} \sim MVN(\boldsymbol{0}, \boldsymbol{T_\pi})$$

$$u_j \sim MVN\left(0, \boldsymbol{T_\beta}\right)$$

The model that included a single polynomial term and treated all coefficients as random at the student and school levels fit best.[4]

By contrast, the CP model expands traditional growth models to include within-year (fall-to-spring) growth components. Given our sample, the CP model allows us to fit between-year growth curve models spanning grades two through six comparable to those in Equations 1-4, as well as fall-to-spring gains for each of those years. In our models, X_{tkij} is the kth CP growth term for time t within student i and school j where

$$y_{tij} = \sum_{k=0}^{5} \pi_{kij}X_{tkij} + e_{tij}. \quad (5)$$

SCHOOL EFFECTIVENESS AND SUMMER LOSS

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

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \qquad (6)$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij}$$
$$\pi_{2ij} = \beta_{20j} + r_{2ij}$$
$$\pi_{3ij} = \beta_{30j} + r_{3ij}$$
$$\pi_{4ij} = \beta_{40j} + r_{4ij}$$
$$\pi_{5ij} = \beta_{50j} + r_{5ij}$$

Finally, the level-3 model for school j is

$$\beta_{00j} = \gamma_{000} + u_{00j} \qquad (7)$$
$$\beta_{10j} = \gamma_{100} + u_{10j}$$
$$\beta_{20j} = \gamma_{200} + u_{20j}$$
$$\beta_{30j} = \gamma_{300} + u_{30j}$$
$$\beta_{40j} = \gamma_{400} + u_{40j}$$
$$\beta_{50j} = \gamma_{500}$$

Though the variance components for this model differ from those under the traditional polynomial, we will nonetheless refer to those variance components as $\boldsymbol{T_\pi}$ and $\boldsymbol{T_\beta}$. As shown in Equation 7, we fit the model in several ways treating different coefficients as both fixed and random, and ultimately found that the model fit best (Bentler, 1990; Fieuws & Verbeke, 2006) when treating all coefficients as random at both the student and school level except for γ_{500}, which is fixed at the school level.

In the CP model, the first three parameters are comparable to those from traditional growth models in Equation 1 that only estimate between-year growth. γ_{000} is the predicted spring score at the centering grade, γ_{100} is the 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 the centering year, γ_{400} is the linear growth rate of change for fall to spring growth, and γ_{500} is the quadratic term for that growth. That is, the first coefficient captures the within year growth for the centering year, and the second coefficient captures the change in that growth rate across years. Thus, the model tells

us not only how much within-year growth occurs in the centering year, but also how we might expect that rate to change as students move through school.

These additional parameters in the CP compared to the traditional growth curve model are straightforward from a modeling perspective, and their ability to produce our estimates of interest may not be immediately apparent. The main reason is that, while the models themselves are not overtly different, the design matrices for the two models are quite different. The specifications of our CP design matrices (X_{ij}) are provided in Appendix A.

To explore the sensitivity of our results to the term at which time is centered, we fit models with time centered at grades 2 and 4.[5] As shown in Table A1, for some models, t corresponds to the test administration in Table 1 such that estimates are centered at second grade. For other models (and as shown in in Table A2), t corresponds to the test administration in Table 1 such that the time variables are centered at fourth grade. Thus, for every research question, we used separate models for math and reading, as well as separate models for the two centering approaches, resulting in a total of four sets of parameter estimates.

Question 1. How much do within-year gains shift over time as students move through school? For this question, γ_{400} is the coefficient of interest. As previously described, this fixed effect is an estimate of school-level changes in fall-to-spring growth over time. For example, for models centered at grade two, a coefficient of -1 would indicate that mean fall-to-spring gains are decreasing linearly at a rate of one RIT point each year between second and sixth grade. Further, γ_{500} describes the quadratic rate of change in fall-to-spring growth over time.

Question 2. How much of the variance in growth is within schools for fall-to-spring versus spring-to-spring estimates? When examining how much of the variance in student test

scores is at the school level for fall-to-spring versus spring-to-spring estimates, the within-school and between-school $\boldsymbol{T}$ and $\boldsymbol{\beta}$ matrices allowed us to produce Intraclass Correlation Coefficients (ICCs). These ICCs show how much of the variance in growth estimates occurs between versus within schools. For example, the variance of the linear school-level estimate for spring-to-spring growth can be divided by the sum of the school- and student-level variance estimates to produce an ICC describing the proportion of the variance in linear spring-to-spring growth at the school level

$$\frac{Var(u_{10j})}{Var(u_{10j})+Var(r_{1ij})} \quad (9)$$

Similarly, the ICC in fall-to-spring growth for the centering grade can be expressed as

$$\frac{Var(u_{30j})}{Var(u_{30j})+Var(r_{3ij})} \quad (10)$$

One problem with the ICCs from our main CP model is that they use only linear parameters from models that include polynomial terms, which complicates their interpretation. Therefore, we also fit the models with only linear growth terms (omitting π_{2ij} and π_{5ij}) and computed the ICCs in Equations 9 and 10 as a point of comparison.

Question 3. How strongly correlated are estimates of school effectiveness that use within- versus between-year estimates? In much of the VAM literature, school effectiveness is compared by producing empirical Bayes estimates of the school-level random effects and correlating them (Loeb, Soland, & Fox, 2014). However, this approach is problematic in the context of our study for a couple of reasons. First, in the CP, there is no single parameter of interest. Just as with the ICCs, comparing only estimates of school contributions to linear spring-to-spring growth (u_{10j}) provides an incomplete picture. Second, the CP model includes parameter estimates of fall-to-spring growth in the first year, and linear and quadratic spring-to-

spring growth based on all years in the sample. Yet, we are mainly interested in comparing fall-to-spring and spring-to-spring growth for a single year, not fall-to-spring growth from a single year and spring-to-spring growth from all years.

Therefore, we take another approach detailed in Thum and Hauser (2015). Specifically, we produce fixed effects estimates, $\hat{\gamma}$, and true parameter variance-covariance matrices for between student ($\widehat{\boldsymbol{T}}_\pi$) and between school ($\widehat{\boldsymbol{T}}_\beta$) variances. We then use those estimates to produce marginal and conditional inferences about school-level achievement and growth. In particular, we use $\hat{\gamma}$ and $\widehat{\boldsymbol{T}}_\beta$ to generate means and standard deviations for distributions by which observed school-level achievement and mean gains in achievement can be converted to Z-scores and compared. We produce these Z-scores in ways that condition growth on initial achievement, an approach not unlike that used in traditional VAM models. Our approach is detailed in Appendix B.

By using a series of contrast matrices (also described in Appendix B), we can produce model-based estimates of a school's mean RIT score for any given time period and gain between any two time periods. We can then standardize the actual mean RIT score gain for a given school between any two time points relative to the model-based distribution of those gains. Thus, we are not reliant on only a single parameter in the model, nor are we restricted to any specific time periods within the data. This approach is much like how student and school achievement and growth norms are constructed (Thum & Hauser, 2015). In simple terms, for a given school's gain G between time points t_1 and t_2, we produce Z scores such that:

$$Z = \frac{G_{t1t2} - \hat{G}_{t1t2}}{SD(\hat{G}_{t1t2})}$$

To make the purpose of the Z scores clearer, consider the downside to comparing between- and within-year growth at the school level using a simple correlation between β_{10j} and β_{30j}. These estimates represent the correlation of only the linear terms in the polynomial comparing within- and between-year growth. Further, that within-year growth is the fall-to-spring gain from year one and the between-year growth is the linear trend across all five years of data. By contrast, our Z-score correlations can be used to look at the spring-to-spring gain for a given year compared to the fall-to-spring gain that occurs within that same spring-to-spring window, and those correlations can be compared for any year in the dataset.

When producing these Z-scores, we compared fall-to-spring gains to spring-to-spring gains when the latter encompassed the former. For example, when centered on 4th grade, we compared spring-to-spring gains between 3rd and 4th grade to fall-to-spring gains during 4th grade. This approach means the time periods being compared overlap. To make these shifting comparisons easier to follow, we present results visually alongside plots of estimated RIT scores over time such that the time periods being compared are clear.

Results

Figure 2 presents plots of our model-based estimates of mean RIT scores in math and reading at each test administration. As the figure demonstrates, the model-based estimates follow a saw-toothed pattern, just as the actual RIT scores did in Figure 1. To confirm the superior fit of the CP relative to the traditional growth curve model for these data, we fit both just as Thum and Hauser (2015) and Thum and Matta (2015) did, and found that the CP fit better based on RMSE, AIC, and BIC statistics while also reducing autocorrelation in the residuals (see Appendix Table A3).

Table 2 presents fixed effects estimates of CP parameters in math and reading, respectively. As previously discussed, results are with time centered at both 2nd and 4th grade for a total of four models. The spring-to-spring math results in column 1 can be interpreted as follows: the intercept suggests that the mean conditional RIT score in spring of second grade was 190.6, the linear growth rate was 14.5 RIT, and that trend decelerated over time per the quadratic term. Meanwhile, the intercept for fall-to-spring growth suggests the mean gain in math during second grade between fall and spring was 13.4 RIT.

Question 1. How Much Do within-year Gains Shift over Time as Students Move through School?

Table 2 also helps answer our first research question about how fall-to-spring gains change as students move through school. Fixed effect estimates indicate that within-year growth in math decelerates linearly at a rate of roughly 1.4 RIT for models centered at grade two and 1.99 RIT for models centered at grade four. For reading, that linear deceleration is almost 4 RIT when centered at second grade and 2.5 when centered at fourth grade. The quadratic term indicates these negative rates of change in fall-to-spring scores slow for reading, but accelerate slightly in math.

These results suggest that fall-to-spring growth slows considerably as students get older. Given the mean gain in math during second grade is 13 RIT, a deceleration of 1.4 RIT suggests mean gains are much smaller in subsequent years as a percentage of that year-one gain. Though more research would be needed to confirm as much, these results indicate that estimated school effectiveness would probably differ dependent on which grades are used in a fall-to-spring model.

Question 2. How Much of the Variance in Growth is within Schools for Fall-to-spring Versus Spring-to-spring Estimates?

Table 3 presents ICCs from the four models, including adaptations of those models that do not include spring-to-spring and fall-to-spring quadratic terms. These ICCs suggest that, across models and subjects, more of the variance in linear growth is at the school level when estimates use within-year rather than between-year growth. Differences are most pronounced when the models do not include quadratic terms. For example, for reading centered at second grade, three times as much of the variance is at the school level for fall-to-spring growth than spring-to-spring. Logically, these results make sense and reflect what has been found in prior literature: more of the variance in student gains are at the school level when those gains are estimated for only the time during the year when students are actually in school (McEachin & Atteberry, 2017).

Question 3. How Strongly Correlated Are Estimates of School Effectiveness that Use within- Versus between-year Estimates?

Figure 3 provides correlations of the conditional Z-scores produced using the approach described in Appendix B (in the figure, these correlations are denoted by "rho"). Again, these are correlations based on a model fit to all five years of data, but use a contrast matrix to compare school-level spring-to-spring gains to fall-to-spring gains in grades three and four. One should note that these estimates can be made for any set of years within the data with the proper adjustment of the contrast matrices.

Results indicate that correlations for between- and within-year growth range from moderate (.399 in reading) to strong (.705 in math centered at 2nd grade). In practical terms, prior research on VAMs suggests that correlations of school effectiveness can have non-

negligible ramifications for rank orderings of schools in a policy context when they fall below .90 (Koedel & Betts, 2010). Using the same threshold, our correlations suggest that estimates of school effectiveness would have differential implications under policies that identify extremely low-performing schools dependent on whether fall-to-spring versus spring-to-spring test scores are used. Unsurprisingly, our results also appear sensitive to the grade used, which could suggest that discrepancies between estimates are more pronounced as students move through school.

Discussion

The passage of ESSA has changed the school accountability landscape in this country significantly. In particular, a large number of states are using school contributions to student academic growth as a component in their accountability systems, oftentimes weighting growth more heavily than static achievement (ESSA Plans, 2018; Klein, 2016). Further, these weighted accountability models will be used under the law to identify and intervene in the lowest-performing schools. Thus, the need for estimates of school effectiveness that are accurate, reliable, and fair is high.

Despite the importance of growth under federal accountability, all state ESSA plans available at the time of our study ignored the seasonality of student test scores, and summer learning loss in particular. This oversight occurred despite research showing that estimates of school effectiveness that ignore summer learning loss produce different rank orderings than those that do not, in part because more of the variance in student gains is attributable to schools when only estimating that growth during the academic year (McEachin & Atteberry, 2017). While many states do not account for summer loss because they only test students in the spring and therefore cannot model within-year growth, there is also a shortage of statistical models that can

be used to compare fall-to-spring and spring-to-spring growth, including estimates of school contributions to that growth.

We begin to address the methodological issues in past research and current policy by applying the CP model in the context of estimating school contributions to student growth. The CP is designed to better describe student growth when trends are markedly seasonal. A major advantage of the CP relative to other models used to estimate teacher and school effectiveness in a way that addresses summer loss is that comparisons of between- and within-year growth can be made from a single model, which helps address sample problems that arise when different students test in fall and spring. Further, trends in within-year growth are directly parameterized in the CP.

Beyond simultaneously modeling fall-to-spring and spring-to-spring growth, the CP is quite different from other models used in the literature on school effectiveness and summer loss. For example, other studies use lag-score VAMs (McEachin & Atteberry, 2017), which do not assume any functional form to student growth and do not rely on an assumption that that there is a vertical scale sufficient to estimate gains. Thus, in the presence of vertically scaled test scores and multiple within-year test administrations, lag-score models may not maximize the usefulness of the data in describing student growth. Further, for the papers that do employ growth rather than lag-score models to estimate school contributions to student growth in the presence of summer loss, the models are fully saturated and therefore do not lend themselves to testing assumptions of model fit, including the functional form of growth (Atteberry & McEachin, 2015; Downey et al., 2008).

Implications for Policy and Evaluation

Besides illustrating the CP for modeling seasonal growth, our results provide evidence on the ramifications of ignoring the seasonality of educational testing data when estimating school effectiveness. To that end, we produce several findings relevant to policymakers, program evaluators, and educators. First, while accountability and evaluation models based on within-year growth tend use fall-to-spring growth to look at school effectiveness (Jensen, Rice, & Soland, 2018), we show that these within-year gains have distinct trends of their own over time. Whereas students in second grade tend to gain about 13-14 RIT during the school year, those gains decelerate at a linear rate of anywhere from 1.4 to 3.9 RIT per year (though there is also a nonlinear component to these rates of change). In practical terms, this finding suggests that fall-to-spring gains are lower as students move through school. Therefore, evaluations of teachers or schools based on fall-to-spring growth are likely sensitive to the grades served by the teacher or school without sufficient controls for those differences in the model. The field would likely benefit from a closer examination of this issue.

Second, we provide further evidence that more of the variance in student achievement and growth is at the school level when using fall-to-spring versus spring-to-spring growth. This result matches what has been found by others (McEachin & Atteberry, 2017). In some cases, twice or three times the variance is at the school level when using fall-to-spring gains. This finding makes intuitive sense: schools are more associated with the academic growth of students during the periods when those students are in school. Though more research is needed in this area, our results are congruous with a theory that schools may be deemed as ineffective under ESSA for academic growth (or lack thereof) that occurs in part when students are not attending school. Given our findings and those of others like McEachin and Atteberry (2017), a broader

policy conversation is likely needed about the extent to which schools should be responsible for how learning changes during the summer months.

We also provide additional evidence that states will identify different schools as effective depending on whether they use fall-to-spring versus spring-to-spring growth. According to our models, while correlations between spring-to-spring and fall-to-spring growth from the same academic year can be highly correlated (.70 or higher), such is not always the case. Further, even correlations in the .70 range would likely lead to practical implications in accountability systems. Koedel and Betts (2010) suggested that correlations of school VAM estimates that fall below .90 can have substantive ramifications for those schools when estimates are used for accountability. Using that standard, none of the zero-order correlations we produce are high enough to indicate that accountability systems rank ordering schools based on fall-to-spring gains will lead to the same conclusions about schools when based on spring-to-spring growth.

While the CP model provides evidence germane to evaluating schools and related policy, there are a few characteristics of the model that should be noted before it is employed. First, our findings can be sensitive to how the CP is parameterized. For example, results often differ when estimates of growth are centered on different grades. On one hand, this sensitivity suggests that, like other models, the CP cannot get us as close as we might like to definitively identifying effective and ineffective schools. On the other, a benefit to the CP model is that it can be used to illustrate the instability of many estimates dependent on which time periods are used in a model and how growth is parameterized.

Relatedly, the CP model is not ideally suited to all policy contexts and research questions. For example, the CP is highly parameterized and involves considerable smoothing across time points. Therefore, in accountability settings where policymakers are only interested in school

contributions to student growth for a single year (or other fairly discrete estimates), the CP model is likely not optimal. Rather, the CP appears ideally suited for comparisons of discrepant time periods, especially those from different years or lengths of time that can often result in inconsistent samples of students being compared. The CP model can also provide estimates of trends in within-year growth across several years, a parameter generally not estimated in other growth models. One should note that the CP includes all the same between-year parameters as the traditional polynomial model, which means the former can do what the latter does, but the CP provides the added benefit of parameterizing aspects of within-year growth.

Limitations of the Current Study

Despite the strengths of our analytic approach, this study has several limitations that bear mention. First, our sample is from one state, therefore results may not generalize to the whole country. Further, while virtually all students in the state took MAP Growth, there may be slight differences between state samples and population. As a check on this potential within-state issue, we estimated models using an entropy balancing (Caliendo & Kopeinig, 2008) weighting scheme based on school-level covariates from federal datasets and found minimal differences in estimates, which is why weighted estimates are not presented.

Second, we relied only on a rough estimate of how much time elapsed between fall and spring test administrations and, therefore, how much instructional time students received. That is, we assumed that students spent three quarters of the year in school, the other quarter out of school. While results were not sensitive to making other generic assumptions about time in school (e.g., students were in school four-fifths of a year instead of three-quarters), these results might shift if more fine-grained calendar data are used. We also did not include winter testing or summer testing data, which could be used to fit within-year polynomial submodels. Our results

should be replicated using more time points within years, and better accounting for the actual amount of instruction students received.

Third, MAP Growth is a low-stakes assessment. In the state we examined, the test is utilized to monitor student progress over the school year and to test goals for growth, but little more. Therefore, one cannot be certain if the same results would hold for high-stakes tests. Despite the low-stakes nature of MAP Growth, there is reason to believe that disengagement among students taking the test is not a primary factor. For example, 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.

Finally, comparable to McEachin and Atteberry (2017), we do not have several important student-level covariates used in many traditional VAMs. For example, we do not have student-level socioeconomic status data. Thus, we cannot be sure how inclusion of such covariates would change estimates.

Future Research

There are several additional applications of the CP model in the context of school effectiveness that are worthwhile to pursue, some of which have been mentioned already. For example, the CP can be re-parameterized such that between-year growth is measured fall to fall rather than spring to spring. Our results could be replicated using the fall-to-fall parameterization to see how different they are. Further, the CP can be parameterized such that within-year growth is estimated directly as spring-to-fall chanes rather than fall-to-spring gains. For researchers primarily interested in estimates of summer learning loss, such models could be used to directly estimate that loss.

Another potential use of the CP model is to compare school contributions to short-term versus long-term growth. For example, one could use the CP to estimate school contributions to student growth during elementary school, and compare those results to fall-to-spring or spring-to-spring estimates from a single year. Such an endeavor is especially relevant to education policy given the increasing focus on how teachers and schools can contribute to the long-term college readiness of students (Conley, 2008). In many ways, accountability policy and the objectives we set for schools are at odds. Whereas state ESSA plans focus on short-term growth over a year or two, other policies and practices tend to emphasize college readiness, which is related to how much students grow academically from Kindergarten through 12th grade. If school contributions to growth in the short- and long-term differ substantively, then policymakers may need to reconsider what the main purpose of schools are and how they should be held accountable for achieving those goals. The CP model could be used to answer related questions without the typical issues of varying samples of students used to estimate short- and long-term growth.

Conclusion

Research shows that schools contributing to fall-to-spring growth in achievement are often not the same ones contributing to spring-to-spring growth, in large part because of learning loss that occurs when students are out of school for the summer (McEachin & Atteberry, 2017). However, researchers and evaluators do not have many statistical tools at their disposal to reconcile differences in estimates of school effectiveness arising from the seasonality of educational testing data (nor do they often have the within-year repeated measures needed to use those statistical tools). In this study, we apply the CP model developed by Thum and Hauser (2015), which can jointly estimate between- and within-year student growth, to fall and spring

testing data over the course of several years. Our results suggest several benefits of using the CP model when data show seasonal trends. We further show that accountability determinations for schools under ESSA are likely sensitive to the time periods used to measure growth, how growth is estimated, and whether models use fall-to-spring versus spring-to-spring gains.

Citations

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

107(2), 238.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational*

Measurement, *50*(2), 204–226.

Briggs, D. C., & Domingue, B. (2012). The gains from vertical scaling. *Applied Measurement in*

Education, 38(6), 551-576.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of*

Educational and Behavioral Statistics, *36*(5), 616–637.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of

propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

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.

Daniel, L. H. (2012). *Comparing cross-classified growth models with and without the cumulative*

effect of teachers to a hierarchical growth model on cross-classified data (PhD Thesis).

University of Pittsburgh.

Downey, D. B., Von Hippel, P. T., & Hughes, M. (2008). Are "failing" schools really failing?

Using seasonal comparison to evaluate school effectiveness. *Sociology of Education*,

81(3), 242–270.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or

measurement model components: Rationale of two-index strategy revisited. *Structural*

Equation Modeling, *12*(3), 343–367.

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

Gershenson, S., & Hayes, M. S. (2018). The implications of summer learning loss for value-added estimates of teacher effectiveness. *Educational Policy*, *32*(1), 55–85.

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.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125–150.

Lockwood, J. R., McCaffrey, D. F., & Savage, C. B. (2017). Errors-in-variables regression: Why Stata's eivreg is wrong and what to do instead. Paper presented at the SREE 2017 Spring Conference, Washington, DC.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, *36*(4), 457-475.

Matta, T., Lathrop, Q., & Thum, Y. M. (2017). apdesign: An implementation of the additive polynomial design matrix. *R Package Version 3.2.2*.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

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

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121-129.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). New York, NY: Sage.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education*, *4*(4), 492–519.

Soland, J. (2017). Is teacher value added a matter of scale? The practical consequences of treating an ordinal scale as interval for estimation of teacher effects. *Applied Measurement in Education*, *30*(1), 52–70.

Thum, Y. M. (2003). Measuring progress toward a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research*, *32*(2), 153–207.

Thum, Y. M., & Soland, J. (2017). School norms for mathematics achievement status, term-to-term growth, and the gender gap. Paper presented at the SREE 2017 Spring Conference, Washington, DC.

Thum, Y. M. (2015b). College readiness benchmarks for grades 5-9 MAP math and reading: Approach and results. Presented at the Annual National Conference on Student Assessment, San Diego, CA.

Thum, Y. M. (2015a). The Effective use of student and school descriptive indicators of learning progress: From the conditional growth index to the learning productivity measurement

system. In R. Lissitz & H. Jiao (Eds.) Value added modeling and growth modeling with particular application to teacher and school effectiveness (237-269). Charlotte, NC: IAP.

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.

von Hippel, P. T., Workman, J., & Downey, D. B. (2017). Are schools (still) a great equalizer? Replicating a summer learning study using better test scores and a new cohort of children.

SCHOOL EFFECTIVENESS AND SUMMER LOSS

Notes

1. In the educational accountability context, a value-added model, or VAM, is simply any statistical model that seeks to isolate the impact of teachers or schools on student academic outcomes, within the constraints of the information available.

2. Growth models for outcomes that are vertically scaled but with scores that do not have equal interval properties are also available. An example is a Gibbons and Bock (1987) model for trends in repeated classifications of patient symptoms. Another example, in economics, is the panel analysis of longitudinal categorical data examined by Hsiao (1986).

3. One should also note that several statistical models have been developed to help manage complications that arise when matching students to teachers or schools, which are related to student mobility. For example, Lockwood, McCaffrey, Mariano, and Setodji (2007) lay out a Bayesian approach that can help account for the methodological and substantive challenges associated with matching students. As discussed later, our own study largely bypasses these issues by assigning students to their modal elementary school. However, such topics have been given ample coverage in the literature, and could be applied to the models in our study.

4. We compared model fit using several approaches. For nested models, we used deviance-based statistics. The greater the drop in the deviance (-2 log likelihood), the more likely the fit is to be significantly better. A likelihood ratio test can then be used to show whether changes in the deviance are significant at a given level. These statistics are available as part of standard output in HLM 7 and are detailed in Raudenbush and Bryk (2002). For all models, we also used the Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and Bayesian information criterion (BIC) (Bentler, 1990; Fieuws & Verbeke, 2006). The AIC and BIC are often used for comparing non-nested models (Bentler, 1990; Fan & Sivo, 2005; Fieuws & Verbeke, 2006). The AIC or BIC for a model is usually written in the form [-2logL +kp], where L is the likelihood function, p is the number of parameters in the model, and k is 2 for AIC and log(n) for BIC. These various tests for nested and non-nested models are the same criteria we used to settle on our preferred CP model.

5. One should note that one need not always recode the time variable in the data then re-run a model in order to ascertain the effect of centering on the estimates. For example, consider the case of simple regression

$$\hat{y} = X_c \hat{\beta}_c$$

where X_c is the design matrix centered at time c and $\hat{\beta}_c$ is the corresponding estimated coefficients. Meanwhile, for time centered at d, we have

$$\hat{y} = X_d \hat{\beta}_d.$$

If one sets the righthand sides of the equations equal

$$X_c\hat{\beta}_c = X_d\hat{\beta}_d,$$

one can solve for $\hat{\beta}_d$ such that

$$\hat{\beta}_d = (X_d'X_d)^{-1}X_d'X_c\hat{\beta}_c \quad ,$$

without necessarily resorting to a re-analysis of the raw data.

Table 1

Analytic Sample Descriptive Statistics

Year	Term	Grade	Test Administration	Race & Gender Proportions				Mean Achievement (RIT Scale)	
				Black	Hisp.	White	Female	Math	Reading
2010	Fall	2	1	0.332	0.076	0.509	0.511	178.170	175.333
2011	Spring	2	2	0.331	0.075	0.512	0.511	191.990	189.367
2011	Fall	3	3	0.331	0.074	0.513	0.510	192.503	190.728
2012	Spring	3	4	0.331	0.073	0.514	0.510	204.730	200.479
2012	Fall	4	5	0.329	0.074	0.514	0.510	203.925	200.299
2013	Spring	4	6	0.328	0.075	0.515	0.510	213.651	207.436
2013	Fall	5	7	0.325	0.075	0.517	0.510	211.804	206.563
2014	Spring	5	8	0.326	0.076	0.516	0.508	220.796	212.596
2014	Fall	6	9	0.333	0.077	0.524	0.511	216.514	211.448
2015	Spring	6	10	0.330	0.079	0.525	0.511	222.030	215.060

Table 2

Fixed Effects Estimates from 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 - spring to spring	190.561	212.428	188.256	206.224
	0.244	0.288	0.267	0.250
2. Linear - spring to spring	14.497	7.370	11.765	6.203
	0.115	0.044	0.088	0.031
3. Quadratic - spring to spring	-1.782	-1.782	-1.390	-1.390
	0.026	0.026	0.021	0.021
4. Intercept - fall to spring gain year one	13.432	10.055	13.745	7.331
	0.123	0.087	0.121	0.061
5. Linear - fall to spring change over time	-1.381	-1.995	-3.953	-2.461
	0.110	0.032	0.104	0.034
6. Quadratic - fall to spring change over time	-0.153	-0.153	0.373	0.373
	0.025	0.025	0.023	0.023

Table 3

Intraclass Correlations by Model and Subject

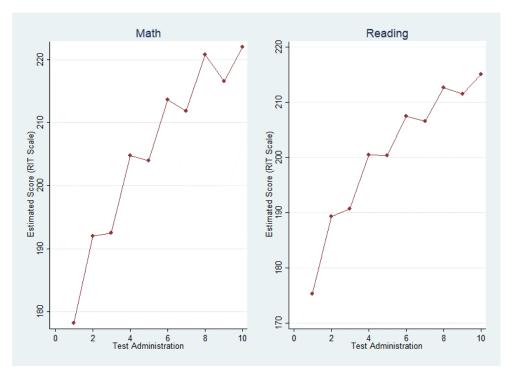|	Math		Reading	
	Centered at 2nd Grade	Centered at 4th Grade	Centered at 2nd Grade	Centered at 4th Grade
Model with no quadratic term				
Spring-to-spring ICC	0.269	0.253	0.152	0.152
Fall-to-spring ICC	0.594	0.565	0.485	0.373
Model with quadratic term				
Spring-to-spring ICC	0.399	0.375	0.178	0.201
Fall-to-spring ICC	0.524	0.499	0.264	0.303

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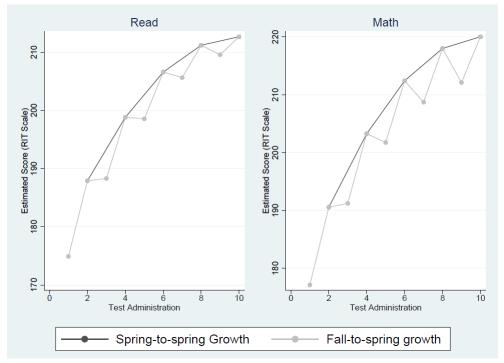

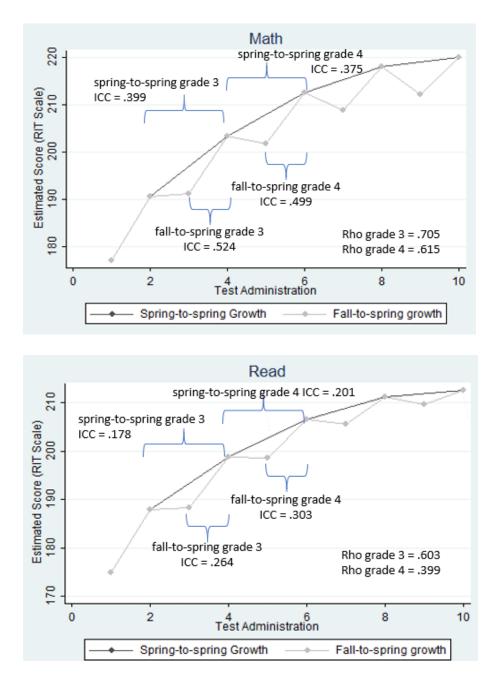*Figure 2. Model-based estimates of mean RIT scores by test administration and subject.*

Figure 3. Model-based estimates of RIT scores with correlations of Z-score estimates and ICCs.

Appendix A: Building a Compound Polynomial Design Matrix

As shown by Thum and Hauser (2015) or Thum (2018), for example, building the CP design matrix begins by specifying within- and between-year design matrices. The first step is to specify the within-year polynomial. Consider the setting in which only two assessments are observed within a year, for example when a score in the fall is followed after d time units by a score in the spring term. If we want to model between-year growth as spring-to-spring, we set the within-year design matrix, $\boldsymbol{D_w}$, equal to

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

where d is an instructional time interval (say, 9/10 of a calendar year) between the fall and spring terms. This design defines the intercept as the predicted spring score and the growth component as the predicted gain from fall to spring. Similarly, when we wish to model fall-to-fall growth between years, $\boldsymbol{D_w}$ is equal to

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

so as to define the within-year growth components as the predicted fall score (intercept term) and the fall-to-spring gain.

For the remainder of this example, we will focus on spring-to-spring between-year growth. Under this version of $\boldsymbol{D_w}$, the first column is a set of intercepts for two within-year time points, fall and spring. The second column represents the time that elapses between fall and spring. In other words, some proportion of a year, d, elapses between fall and spring. (One should note that d can eventually be replaced with specific instructional calendar data if available.)

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

$$\mathbf{CP1} = \mathbf{G} \otimes \mathbf{D_w} =$$

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

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

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. For example, the values in the first two columns and in the first row represent fall of 2010, and the values in the first two columns and the second row represent spring of calendar year 2011. Similarly, the values in the last two columns in the last row represent spring of 2015.

Growth or change in the predicted spring score and the fall-to-spring gain for each year may then be described by a second between-year polynomial. This second design matrix for between year (spring to spring) growth, $\mathbf{D_b}$, is identical to the design matrix for a traditional growth model. For example, the between-year spring-to-spring scores and linear trend in fall-to-spring gains may *each* be described, for example, by $\mathbf{D_b} =$

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

In 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.

We then produce our second CP matrix, **CP2**, using the following Kronecker product with our between-year design matrix, D_b:

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

This function produces the following 10 x 6 matrix, **CP2**:

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

CP2 can be thought of as the fall status, linear slope, and quadratic slope over grades.

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

$$\mathbf{CP} = \mathbf{CP1} * \mathbf{CP2} =$$

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

In this matrix, the first three columns represent the intercept, linear growth, and quadratic growth

terms for the between-year spring-to-spring component. Similarly, columns four through six

represent the intercept, linear, and quadratic growth terms across years for the within-year fall-

to-spring gains. The final CP design matrix centered at grade two, including year, term, grade,

and test administration, can be found in Table A1 below. Similarly, A2 presents the same CP

design matrix but centered at grade four. The model can now be fit as in Equations 5-7, with the

first five columns in matrix **CP** corresponding to each of the six π parameters.

Appendix B. 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:

$$C = \begin{pmatrix} c_1' \\ c_2' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \quad \text{(B1)}.$$

Using this contrast matrix, we can produce

$$CY = \begin{pmatrix} Y_t \\ Y_{t+1} - Y_t \end{pmatrix} = \begin{pmatrix} Y_t \\ G \end{pmatrix} = P \quad \text{(B2)}$$

where Y_t is the starting RIT and G is the gain. The predicted achievement is $\hat{y} = c_1' A \hat{\gamma}$ and the

marginal predicted gain between times 1 and 2 then becomes

$$\hat{G} = c_2' A \hat{\gamma} \quad \text{(B3)}$$

where $\hat{\gamma}$ is the 1×6 vector of school fixed effects estimates from Equation 7 and A is a matrix

composed of the rows from the appropriate design matrix corresponding to the time points of

interest (see Tables A1 and A2). For example, when looking at growth between fall and spring

of 2nd grade using spring-to-spring between-year growth centered at grade two, A would

correspond to the first two rows of the design matrix in Table A1. Conventional results for

expectations of random variables give the standard error of $\hat{G}$ as

$$se(\hat{G}) = \sqrt{c_2' A Var(\hat{\gamma}) A' c_2} \quad \text{(B4)}.$$

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

$$V_s = C\left[A Var(\hat{\gamma}) A' + A H_3 \widehat{T}_\beta H_3' A'\right] C' \quad \text{(B5)}.$$

Here, H_3 is a selection matrix that identifies the random coefficients among β_j for schools with

estimated variance-covariances of $\widehat{T}_\beta$. As an example, if a model included six fixed effects but

only five random effects at the school level (as ours does), H_3 selects from T_π (see Equation 7)

the terms corresponding to the five parameters with random effects.

From here, one can take the square root of 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

$$G_{Sj}^* = \hat{G} + V_{s[2,1]} \cdot V_{s[1,1]}^{-1} \cdot \left(\bar{y}_{j1} - \hat{y} \right) \quad \text{(B6)}$$

And the expected conditional standard deviation for those gains as

$$SD_{Sj}^* = \sqrt{V_{s[2,2]} - V_{s[2,1]} \cdot V_{s[1,1]}^{-1} \cdot V_{s[1,2]}} \quad \text{(B7)}.$$

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

This approach need not be limited to only two time points, nor to adjacent points in time. The contrast matrix C can be adjusted to support comparisons of gains between multiple time points. For example, to explore the relationships amongst within- and between-year gains in our study, we used the following contrast matrix

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{(B8)}.$$

SCHOOL EFFECTIVENESS AND SUMMER LOSS

In conjunction with a Y_i matrix consisting of the observed RIT scores from spring of second

grade, fall of third grade, and spring of third grade, the contrast matrix produces:

$$C * Y_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix} * \begin{pmatrix} RIT_{spring2nd} \\ RIT_{fall3rd} \\ RIT_{spring3rd} \end{pmatrix} = \begin{pmatrix} RIT_{spring2nd} \\ RIT_{spring3rd} - RIT_{fall3rd} \\ RIT_{spring3rd} - RIT_{spring2nd} \end{pmatrix} \quad (B9)$$

$$= \begin{pmatrix} Initial\ RIT \\ Fall - Spring\ Gain \\ Spring - Spring\ Gain \end{pmatrix}$$

Z-scores can then be produced for the initial RIT, fall-to-spring gain, and spring-to-spring gain.

This is the contrast matrix we used when comparing within- versus between-year contributions

to student growth.

Table A1

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 A2

Spring to Spring Cumulative Polynomial Design Matrix Centered at Grade 4

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	-2	4	-1	2	-4
2011	Spring	2	2	1	-2	4	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	0	0	-1	0	0
2013	Spring	4	6	1	0	0	0	0	0
2013	Fall	5	7	1	1	1	-1	-1	-1
2014	Spring	5	8	1	1	1	0	0	0
2014	Fall	6	9	1	2	4	-1	-2	-4
2015	Spring	6	10	1	2	4	0	0	0

Table A3

Comparing the Fit of the Regular and Cumulative Polynomial

Metric	Math		Reading	
	Polynomial	CP	Polynomial	CP
RMSE	4.706	3.550	4.989	4.292
AIC	2665859.454	2573840.794	2724483.903	2681070.466
BIC	2666011.466	2574014.522	2724635.801	2681244.064

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.