ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-60

Effect Sizes for Measuring Student and School Growth in Achievement: In Search of Practical Significance

James Soland
University of Virginia

Yeow Meng Thum
NWEA

Effect sizes in the Cohen's d family are often used in education to compare estimates across studies, measures, and sample sizes. For example, effect sizes are used to compare gains in achievement students make over time, either in pre- and post-treatment studies or in the absence of intervention, such as when estimating achievement gaps. However, despite extensive research dating back to the paired t-test literature showing that such growth effect sizes should account for within-person correlations of scores over time, such achievement gains are often standardized relative to the standard deviation from a single timepoint or two timepoints pooled. Such a tendency likely occurs in part because there are not many large datasets from which a distribution of student- or school-level gains can be derived. In this study, we present a novel model for estimating student growth in conjunction with a national dataset to show that effect size estimates for student and school growth are often quite different when standardized relative to a distribution of gains rather than static achievement. In particular, we provide nationally representative empirical benchmarks for student achievement and gains, including for male-female gaps in those gains, and examine the sensitivity of those effect sizes to how they are standardized. Our results suggest that effect sizes scaled relative to a distribution of gains are less likely to understate the effects of interventions over time, and that resultant effect sizes often more closely match the estimand of interest for most practice, policy, and evaluation questions.

VERSION: May 2019

Suggested citation: Soland, J. & Thum, Y.M. (2019). Effect Sizes for Measuring Student and School Growth in Achievement: In Search of Practical Significance (EdWorkingPaper No.19-60). Retrieved from Annenberg Institute at Brown University: http://edworkingpapers.com/ai19-60

Effect Sizes for Measuring Student and School Growth in Achievement:

In Search of Practical Significance

James Soland
Assistant Professor – University of Virginia
Associated Research Fellow – NWEA
(Corresponding Author)

Yeow Meng Thum
Senior Research Fellow – NWEA

University of Virginia
Curry School of Education
405 Emmet Street
Charlottesville, VA 22904

NWEA
121 N.W. Everett Street
Portland, OR 97209
Ph. (503) 444-6449
jim.soland@nwea.org
(Please contact James here)

May 17, 2019

Abstract

Effect sizes in the Cohen's *d* family are often used in education to compare estimates across studies, measures, and sample sizes. For example, effect sizes are used to compare gains in achievement students make over time, either in pre- and post-treatment studies or in the absence of intervention, such as when estimating achievement gaps. However, despite extensive research dating back to the paired t-test literature showing that such growth effect sizes should account for within-person correlations of scores over time, such achievement gains are often standardized relative to the standard deviation from a single timepoint or two timepoints pooled. Such a tendency likely occurs in part because there are not many large datasets from which a distribution of student- or school-level gains can be derived. In this study, we present a novel model for estimating student growth in conjunction with a national dataset to show that effect size estimates for student and school growth are often quite different when standardized relative to a distribution of gains rather than static achievement. In particular, we provide nationally representative empirical benchmarks for student achievement and gains, including for male-female gaps in those gains, and examine the sensitivity of those effect sizes to how they are standardized. Our results suggest that effect sizes scaled relative to a distribution of gains are less likely to understate the effects of interventions over time, and that resultant effect sizes often more closely match the estimand of interest for most practice, policy, and evaluation questions.

Keywords: effect sizes, norms, student achievement, student growth, standardization.

Effect Sizes for Measuring Student and School Growth in Achievement: In Search of Practical Significance

Effect sizes are a metric used frequently in education program evaluation and policy research. For example, effect sizes have been reported to describe estimates of teacher effectiveness (Soland, 2017), school effectiveness (Ladd & Walsh, 2002), achievement gaps (Quinn, Cooc, McIntyre, & Gomez, 2016), seasonal learning patterns (von Hippel, Workman, & Downey, 2018), and educational interventions (Bloom, 1988). Standardizing such evaluation and policy estimates is popular because the estimates can be compared more easily across studies and standard deviation (SD) units are often more familiar than scale score units from achievement tests. Despite the prevalence of effect sizes, their complexities and shortcomings were somewhat infrequently discussed, at least until a series of studies beginning with Hill, Bloom, Black, and Lipsey (2008) and Bloom, Hill, Black, and Lipsey (2008). These complexities often relate to assessing the practical significance of an effect.

Many such complexities arise because one cannot always be clear whether the variance used in the effect size denominator ideally matches the sample used in the numerator for the desired inference. That is, if effect sizes should not only be about standardization, but also producing a metric relevant to the practice or policy question of interest (Kelley & Preacher, 2012), then what denominator produces an effect size most germane to the estimand of interest? Examples of the difficulties in selecting the most meaningful denominator relative to the numerator abound. For example, Bloom et al. (2008) discuss cases where school-level achievement is standardized relative to the student-level test score distributions. While such a decision may be justified mathematically and interpretively, one could also imagine such a

choice leading to under-estimates of school gains and losses given school-level variances are generally smaller than student-level variances.

Another form of complexity in pairing a numerator to the appropriate denominator relates to producing effect sizes for student growth. As Lipsey et al. (2012) pointed out, these issues of practical significance as they relate to student growth over time merit additional discussion and empirical evidence. Virtually all the studies we are aware of standardize growth relative to achievement at a point in time or to the pooled standard deviation at the two timepoints being compared, including thoughtful research on effect sizes conducted by Bloom et al. (2008), Hill et al. (2008), and Lipsey et al. (2012). There are also numerous examples in applied research beyond the effect size literature (e.g., Quinn et al., 2016; von Hippel et al., 2018). However, little discussion is given to the fact that standard deviations that are clearly quite different are often pooled, or that most effect sizes for gains ignore correlations between pre- and post-tests.

On one hand, there is nothing wrong as a matter of procedure with such approaches to producing effect sizes for growth in achievement. On the other, one could argue that an effect size more relevant to most policy and evaluation questions about growth would involve standardizing relative to a distribution of gains rather than a distribution of achievement. Such an approach would mean that program evaluators could compare, say, the mean gain for a school assessed pre- and post-treatment with norms for student growth in the population over the period of intervention. The effect size would then tell evaluators how much growth the treatment school experienced relative to typical growth in the population. By comparison, standardizing the gains for that school relative to the pooled SD for the two timepoints means the effect size does not account for a correlation between achievement at Time 1 and Time 2, and therefore is comparing the intervention to the expected range of scores absent growth. Leaving aside these

interpretive complexities, one could imagine that standardizing gains relative to the pooled SD of the test scores from the two timepoints might understate the effect of the intervention given the SD of the gains is smaller than the pooled SD when the pre- and post-test scores are positively correlated.

Such decisions about how to produce effect sizes for growth also have implications for some of the most common policy metrics used in education, including quantifying changes in achievement gaps over time. In order to tell if an effect size for an achievement gap or the change in that gap is large for a given sample, researchers have suggested comparing effect sizes to normative trends in gaps and changes in gaps (Bloom et al., 2008; Hill et al., 2008; Lipsey et al., 2012). For example, Bloom et al. (2008) used a national sample of test scores to examine male-female, black-white, Hispanic-white, and high- versus low-socioeconomic status gaps, and set empirical benchmarks for those gaps. Those empirical benchmarks were intended as a point of comparison for gap effect sizes found in other studies. However, while those empirical benchmarks are indeed valuable, Bloom et al.'s longitudinal data relied primarily on a small subset of districts, which meant that the distributions of gains were not nationally representative. Further, all effect sizes for gains were standardized relative to the pooled SD of test scores from two timepoints, not to the distribution of test score gains. A principal reason for their approach is likely that no suitable national distributions of achievement gains existed.

Our study exploits data from a national sample of schools (weighted to better match the nationwide population of US public schools) whose students took the Measures of Academic Progress (MAP) Growth, a cross-grade computerized adaptive achievement test with a vertical scale. We also present a novel multilevel model that can be used to simultaneously estimate between and within year growth. Using this model, effect sizes for growth for any between- or

within-year period of time in the sample can be produced in such a way to account for days of instruction, initial achievement level, summer loss, and other relevant factors. This data-model combination means we were able to examine the practical consequences of taking different approaches to producing effect sizes for test score gains, as well as provide more comprehensive empirical benchmarks for changes in certain achievement gaps over time. Specifically, we investigated three research questions:

1. How different are effect size estimates of growth in student achievement when standardized relative to a distribution of gains rather than a distribution of scores from one timepoint (or pooled SDs from two timepoints)?
2. How different are effect size estimates of school-level growth when standardized relative to distributions of student- versus school-level gains?
3. What are national benchmarks for growth in male-female achievement gaps over time, and how sensitive are those empirical benchmarks to the variance used in the denominator?

We begin the study by reviewing relevant literature, then using that literature to present a taxonomy of effect sizes for student growth. Using results from the 2015 MAP Growth norms, we then explore our three research questions to illustrate how much growth effect sizes differ dependent on the distribution used to produce the SD followed by estimates of male-female achievement gaps using national norms for student growth. Finally, we discuss what our results mean for estimating effect sizes, particularly related to student achievement growth. Altogether, our analyses address an overarching question: how can we best evaluate the practical significance of learning gains that students make relative to a practice or policy estimand?

Background

In our paper, we are mainly interested in effect sizes for gains in achievement test scores over time, including comparing gains across groups. The need for such research has been articulated often, including recently by Baird and Pane (2019), who grappled with the issue of practical significance of student achievement gains by trying to understand how to quantify years of learning. However, we begin by providing a broader review of the literature in order to define "effect size" and consider research on how the unit of standardization (denominator) might affect inferences based on effect sizes, including at a single point in time.

Research on Effect Sizes and Their Practical Significance

While effect sizes in the Cohen's *d* family (1977,1988) tend to be thought of as a mean or mean difference divided by a standard deviation in order to produce a standardized estimate of an effect, Kelley and Preacher (2012) remind us that an "effect size" could be much more broadly understood as "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (p. 137). Accordingly, an effect size is not merely about standardization; it cannot be extricated or separated from a given question of interest. In fact, Kelley and Preacher (2012) argue that an effect size need not subtract off the mean in the numerator if the units being compared already have some intuitive or intrinsic meaning. Thus, an effect size is an attempt to illuminate practical significance.

With Kelley and Preacher's (2012) framework in mind, a range of studies have been conducted on using effect sizes to determine the practical significance of various educational outcomes. Depending on the design of the study, and the analyses required, a large number of effect sizes have been proposed (Rosenthal & Rosnow, 1991). Oftentimes, studies revert to the "rules of thumb" proposed by Cohen (1977, 1988), who dubbed 0.2 as "small," 0.5 as "medium," and 0.8 standard deviations as "large" when comparing the difference between the means of two

groups. One should note that correlations and odds ratios can also be thought of as other classes of effect sizes, and similar rules of thumb have been developed for them. However, such considerations, though worthy of additional conversation in a growth context, are beyond the scope of our study.

Nevertheless, as Lipsey et al. (2012) pointed out, comparing educational effect sizes based on the generic cut points described by Cohen (1988) is "like characterizing a child's height as small, medium, or large, not by reference to the distribution of values for children of similar age and gender, but by reference to a distribution for all vertebrate mammals" (p. 4). Said differently, Cohen's empirical benchmarks do not account for the specific outcome variables, interventions, and participant samples in a given study—to the question of interest. In fairness to Cohen, he recognized such limitations in the original work (1977,1988), but those caveats are often not acknowledged in research. Regardless, more refined, context-specific empirical benchmarks are needed (Bloom et al., 2008; Hill et al., 2008).

A handful of studies have laid out approaches for developing meaningful empirical benchmarks against which to compare effect sizes specifically related to student growth in achievement. Hill et al. (2008) articulated several ways to benchmark the magnitude of effect sizes to establish the practical significance of those effects. One type of benchmark is to compare student growth to normative expectations for change. That is, growth can be compared to expectations for growth or change in a "typical" setting or school. In practical terms, one could compare growth for students in a sample of interest to normative expectations for growth during the school year. Hill et al. (2008) illustrate this point by estimating effect sizes for cross-sectional differences in mean test scores across kindergarten through 12^{th} grade based on national norming samples of seven separate standardized tests. In so doing, they show that a practically

significant effect size for growth observed in a given sample is dependent on the grade being studied because mean differences in test scores across adjacent grades differ substantively dependent on the grades being compared. One should note that their decision to use cross-sectional student gains stemmed largely from a dearth of available longitudinal data that are nationally representative.

As another variety of empirical benchmarks, Hill et al. (2008) describe policy relevant points of comparison. Specifically, when examining achievement gaps, they mention that gap effect sizes should be compared to existing differences by race and biological sex, especially when those differences can be based on relevant norms. For example, an effect size of 0.10 may represent a smaller substantive change for some achievement gaps (e.g., for Black-White) than for others (e.g., male-female). Such findings help reinforce the point made by Kelley and Preacher (2012) that effect sizes should be tied specifically to the research question being asked. In related work, Lipsey et al. (2012) provided additional evidence to support the empirical benchmarks identified by Hill et al. (2008) and articulated the importance of norms, stating that effect sizes should be compared to those from relevant norms, and that those norms should be "based on distributions of effect sizes for comparable outcome measures from comparable interventions targeted on comparable samples" (p. 4).

Research, including studies specific to student achievement, growth, and gaps, have used Hill et al.'s framework to produce empirical benchmarks. Most relevant to our own study, Bloom et al. (2008) used data from their national standardized tests to produce norms for growth effect sizes by grade and achievement gaps by race and biological sex. Through those analyses, they showed that gains in the early elementary grades are followed by gradually declining gains in later grades, and that student achievement gaps are relatively small for biological sex and

much larger by race. Thus, applying generic effect size cutoffs for such analyses will likely produce misleading results. As importantly, they compared effect sizes for growth based on cross-sectional versus longitudinal data, and show that the effect sizes differ.

While all of these studies cited call for more thoughtfulness about using empirical benchmarks to determine practical significance, they simultaneously pay less attention to the variances used to determine the effect sizes in the first place. That is, focus is given to comparing effect sizes to those from a relevant normative distribution, but little discussion beyond what is provided by Bloom et al. (2008) is given to the variances used to produce the effect sizes. Thus, they leave two broad gaps in the literature. First, there is little empirical evidence on how much the pairing between the estimand of interest (i.e., in the numerator) and the variance used in the denominator could affect results, especially related to student gains or growth. Relatedly, little is known about how consequential choices about the variance used to standardize the effects actually are. Second, the empirical benchmarks they provide for growth generally do not include longitudinal data beyond limited samples, and are typically standardized relative to scores at a given timepoint, not a distribution of gains. Thus, their benchmarks for growth are incomplete and may not ideally match the estimand, as we discuss below.

A Closer Look at Effect Sizes Used to Compare Student Test Score Gains

Before turning to our specific analyses, consideration of how Cohen's *d* effect sizes for test score gains are currently generated in the literature is warranted. (While there are many versions and reworkings of Cohen's *d*, including Hedges's *g*, we do not discuss them in detail here [Hedges & Hedberg, 2007].) As a point of reference, Table 1 present a generic cross-tabulation of school grade and time period being compared. Under a cross-sectional design, one could examine mean differences in test scores for adjacent grades at a single point in time. Such

an approach is taken by Hill et al. (2008). While the cross-sectional design does not actually involve test score gains, such an analysis was used by Hill et al. (2008) in part due to an absence of nationally representative longitudinal test score data. Under a longitudinal design, one could follow a cohort of students as they move through the grades over time. This approach was taken by Bloom et al. (2008) using a limited sample of district test score data.

INSERT TABLE 1

Effect sizes for gains made by a single group. When discussing test score effect sizes, we will examine the mean test score for grade g, $\bar{y}_g$, and the variance of that test score for a given sample and grade, s_g^2. Using a cross-sectional set of students (e.g., comparing students who are in grade four to those in grade three at Time 1 in Table 1), Hill et al. (2008) produced the following effect size (we use grades three and four as an example for clarity):

$$\frac{\bar{y}_4 - \bar{y}_3}{\sqrt{\frac{s_4^2 + s_3^2}{2}}} \quad (1)$$

This effect size can be interpreted as the mean cross-sectional test score difference between 3rd and 4th grade students at Time 1 scaled relative to the pooled SD of the scores in those grades. One should note that we begin by presenting a cross-sectional effect size even though we are only interested in effect sizes for gains because Hill et al.'s intent was to better understand how achievement changes over time. However, due to their sample, they were not in a position to examine longitudinal data. Therefore, their approach is an extreme example of the potential mismatch between the estimand of interest (growth in achievement over time) and the estimator used.

By contrast, Bloom et al. (2008) produced an effect size using Equation 1, but for longitudinal gains (comparing students in grade four at Time 2 to those in grade 3 at Time 1).

This effect size presents the difference in mean test scores taken between two timepoints for a longitudinal cohort of students in units of the pooled SD of the scores at those two grades and timepoints. Thus, they report the mean difference in achievement between two timepoints as students move through school in units of the pooled SD of scores at those timepoints.

An oft overlooked problem with Bloom et al.'s (2008) longitudinal effect size is that it ignores the correlation between within-student test scores over time. By contrast, Equation 2 below accounts for that correlation:

$$\frac{\bar{y}_4 - \bar{y}_3}{\sqrt{s_4^2 + s_3^2 - 2cov(y_4, y_3)}} \quad (2)$$

Here, $cov(y_4, y_3)$ is the covariance of the test scores between Grade 4 at Time 2 and Grade 3 at Time 1. Without including this term, the assumption is that scores are uncorrelated over time (Gibbons, Hedeker, & Davis, 1993; Zimmerman, 1997). Research suggests, however, that this assumption is not tenable. For example, studies show that student achievement is quite trait-like over time (Soland & Kuhfeld, 2019) and that correlations of test scores can exceed .70. Further, when using longitudinal data, the effect size in Equation 1 is likely to be smaller than the effect size in Equation 2 because the denominator for the latter will typically be smaller when the correlation between within-cohort test scores across time points is subtracted off the pooled variance. While Equation 2 is somewhat foreign in the educational effect size literature, it is common practice in the general hypothesis testing literature for paired t-tests, which roughly follow Equation 2 but use the standard error in the denominator (Gibbons, Hedeker, & Davis, 1993; Zimmerman, 1997).

Another approach that builds on Equation 2 is to compare gains for a given student or groups of students to norms for "typical" growth in the population. Such an effect size would help answer the question: how much did, say, a particular set of students who underwent an

intervention grow relative to typical growth over the same time period? For example, one could estimate the following effect size

$$\frac{(y_4 - y_3) - \mathbb{E}(y_4 - y_3)}{s_{(y_4 - y_3)}} \quad (3)$$

Where $(y_4 - y_3)$ is the observed gain for some student or group of students, $\mathbb{E}(y_4 - y_3)$ is the mean gain in the population, and $s_{(y_4 - y_3)}$ is the SD of the gain. Using the SD of the gain implicitly accounts for the correlation between test scores at two timepoints.

Having a distribution of gains also means that effect sizes can be produced that are relevant to understanding how growth changes as students move through school. For instance, one might be interested in the quantity:

$$\frac{\mathbb{E}(y_5 - y_4) - \mathbb{E}(y_4 - y_3)}{s_{(y_4 - y_3)}} \quad (4)$$

which would be interpreted as how much the mean gain in test scores between 4th and 5th grade differed from the mean gain between 3rd and 4th in units of SDs of the 3rd to 4th gain. Equation 4 is a thus a tool for understanding difference-in-differences. In practical terms, it tells us how much growth in one grade is increasing or decreasing relative to growth in the prior year. Bloom et al. (2008) and Hill et al. (2008) both attempted to quantify that same change in achievement over time, but did so using an effect size that does not account for the correlation between test scores at different timepoints, nor using the prior year's growth as a baseline. Thus, if the estimand is how much student growth is increasing or decreasing year over year, we would argue Equation 4 comes closer to matching that estimand.

Effect sizes for gains made by two different groups. The effect sizes for a single group of students followed over time can be extended to contrasts between gains made by two groups.

Given the empirical example we use, we will compare gains in achievement for males, m, versus females, f. A commonly used effect size for such a purpose is

$$\frac{(\bar{y}_{4m} - \bar{y}_{3m}) - (\bar{y}_{4f} - \bar{y}_{3f})}{\sqrt{\frac{s_{3m}^2 + s_{3f}^2}{2}}} \quad (5)$$

Here, $\bar{y}_{4m}$ is the 4th grade mean for males, and s_{3m}^2 is the SD of achievement test scores for males in 3rd grade. Like Equation 1, Equation 5 does not account for within-person correlations between gains at two timepoints. In practical terms, the denominator essentially assumes there is no growth occurring.

As an alternative, one could fit an effect size that does account for the correlation of test scores over time that is comparable to Equation 4:

$$\frac{(\bar{y}_{4m} - \bar{y}_{3m}) - (\bar{y}_{4f} - \bar{y}_{3f})}{\sqrt{\frac{s_{(4m-3m)}^2 + s_{(4f-3f)}^2}{2}}} \quad (6)$$

In Equation 6, $s_{(4m-3m)}^2$ is the variance of the achievement test score *gain* between grades three and four for males. Thus, the denominator is the pooled SD of the gains made by males and females between the timepoints. Because the gain implicitly accounts for the correlation in test scores over time, and the variance in gains is likely to be smaller than the variance in test scores at a given point in time, Equation 5 is likely to understate the male-female gap in gains.

Effect sizes for gains made by schools. In program evaluation, one might be especially interested in making inferences about aggregated gains at the school level rather than student-level gains. Therefore, one might prefer an effect size scaled relative not to a distribution of student-level gains, but of school-level gains. That is, one could adapt Equations 3-4 to use

gains and SDs at the school rather than student level. Such an effect size would tell us how much the mean gain in a school (possibly pre- and post-intervention) relates to typical school-level growth. As previously discussed, using a denominator at the student level could lead to under-estimates of strong school performance and over-estimates of weak school performance given school-level variances are generally smaller than student-level variances. This phenomenon is illustrated in a stylized scatterplot for a random sample of schools in our data in Figure 1. As the figure shows, the variance of both pre-test achievement and test score gains are much smaller at the school level.

As is hopefully clear in this brief review of the literature, we are not the first to suggest that effect sizes comparing changes between time points should account for the associated correlation (e.g., Zimmerman, 1997), nor that the most appropriate variance to place in the denominator is the SD of the gains (Gibbons et al., 1993). Many fewer studies make the related suggestion that gains for a given sample can be compared to typical gains in the population, and standardized relative to the SD of the gains. Yet, these options are discussed infrequently in the educational effect size literature. One likely reason is that few results for longitudinal, vertically scaled assessments are available for such purposes.

Creating a taxonomy of effect sizes for growth. To make the connections among Equations 1-6 clearer (and, thereby, the differences in the effect sizes for student growth), Table 2 presents a taxonomy of effect sizes for estimates of growth in achievement. As the table shows, one should first determine whether the effect size is for a single group versus a comparison of two groups, and whether those groups are at the student or school level. Next, there may be a determination of whether the effect size is ideally intended to describe growth for a sample or the population. While the idea of an effect size for the population may not be

immediately intuitive, this notion often underlies the idea of empirical benchmarks. Most notably, Hill et al. (2008) and Bloom et al. (2008) are primarily trying to create effect sizes for achievement gains in the population against which to compare gains from a given sample, study, or intervention (benchmarks). For example, a district administrator may wish to evaluate a specific program, in which case the intent is to compare pre- and post-test scores for the participating students to the population gains described by Hill et al. (2008) and Bloom et al. (2008).

INSERT TABLE 2

After making decisions about the number of groups, unit of analysis (student or school), and whether using a sample or some representation of the population, one should consider the specific research question of interest. In some cases, the research questions are different but equally valid. In others, the justification for one research question over another is unclear. For example, we would argue an effect size for student gains that ignores the within-student or within-school correlation of test scores over time (Equation 1 and Equation 4) is not especially justifiable relative to effect sizes that do account for the correlation (Equations 2, 3, and 5). However, among effect sizes that do account for the correlation, all inquiries are arguably valid. For instance, the difference between Equations 2 and 3 is that one standardizes the gain, but the other standardizes the gain relative to expected growth in the population. That is, Equation 2 quantifies the gain, whereas Equation 3 tells us how much a student or school gained (perhaps in the presence of an intervention) relative to "normal" or "typical" growth. Thus, Equation 2 examines a difference whereas Equation 3 examines a difference-in-differences.

In this paper, we provide empirical benchmarks for growth that are useful for determining the practical significance of research findings on achievement and growth for students and

schools in mathematics and reading assessments over time. Specifically, we estimate effect sizes using the different approaches provided in Table 2, then compare them quantitatively and interpretively. We conclude with a discussion meant to help chart a sensible path for generating effect sizes for growth moving forward.

Methods

In this section, we describe the 2015 MAP Growth norming results that we used for our empirical analysis. We estimate growth based on a relationship between achievement and instructional exposure. The proxy for instructional exposure is the number of days of instruction a student has had between achievement tests adjusted for variation in school calendars. We leverage that relationship between instructional time and achievement to define an effect size in terms of number of weeks of exposure.

Sample

This study employs MAP Growth results from six age-cohorts of students who took the MAP Growth mathematics and reading tests in the 2011-12, 2012-13, and 2013-14 school years. Up to three years of longitudinal test scores are used for each student in an age-cohort, with a maximum of nine scores spanning three grade levels over the period of three school years. Age-Cohort 1 consists of a group of students who attended grade 3 in the 2012-13 school year. Similarly, Age-Cohort 6 consists of a group of students who attended grade 8 in the 2012-13 school year. In each age-cohort, about 150,000-200,000 students attending some 1,400 schools contribute well over 500,000 scores to each analysis (see Table 3).

INSERT TABLE 3

Schools for each study cohort were sampled from a larger database of schools that administer MAP Growth. To achieve meaningful generalizability to portray student and school

performance on MAP among the US population of public schools, we derived school-level post-stratification weights. The NCES school characteristics included measures of school poverty, racial make-up, type (e.g., charter), grades served, and location. Details on the weights and weighting procedure are provided in the Appendix.

Measures of Achievement

Students in the sample took NWEA's MAP Growth assessment in reading and mathematics. These assessments are typically administered three times a year in the fall, winter, and spring. MAP Growth is tied to local standards in each state and is computer adaptive with the ability to adapt above and below the content for a given student's grade-level. Each test takes approximately 40 to 60 minutes depending on the grade and subject area. Students respond to assessment items in order (without the ability to return to previous items), and a test event is finished when a student completes all the test items (typically 40 items for reading). Test scores, called "RITs," are reported in an IRT-based metric. Further, the test is vertically scaled, allowing for arithmetic comparisons in evaluating growth across grades. In general, MAP Growth is not used for high-stakes purposes, though some districts and schools may use it to help screen students for special education and gifted programs.

Analytic Approach

Instructional time. An important feature of the data that is exploited in our effort to describe change in achievement over time is the variation in the student's testing schedule across classrooms and schools, due largely to various administrative considerations of a school's calendar and to the availability of testing laboratories. Figure 2 displays the data available for the analysis. Notice the substantial variation the schedule of tests within each term or season. Rather than ignore or suppress such variation, we converted each student's location on the testing

schedule into a proxy for the amount of instruction she has received by determining the calendar date on which instruction began at the school. An advantage of the strong variation in student testing schedules is that the variation in the predictor variable improves the description of the relationship between achievement over time. Interpolations of this relationship between time points will be more accurate as a result.

Joint distributions of predicted scores are obtained from parameter estimates of the fitted multilevel growth model. Achievement norms may thus be derived for any point of the instructional calendar of a grade level (e.g., 180 instructional days for a typical 180-day school calendar). Similarly, marginal gains between any two points on the instructional calendar, within or between adjacent grade levels, may be determined. Growth or gains that are conditional on initial performance are also estimated from direct manipulation of the joint distributions of predicted scores. We describe that growth model below.

Modeling growth. Specifically, we denote the score received by student i in school j at test term t by $t = 1,2, \dots, n_{ij}$ by Y_{tij}. We describe the score trends of students and schools with the generic three-level hierarchical linear model

$$\text{Within Student: } Y_{tij} = \boldsymbol{a}'_{tij}\tau_{ij} + e_{tij} \quad (7)$$

$$\text{Between Students: } \tau_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}_j + \boldsymbol{r}_{ij}$$

$$\text{Between Schools: } \boldsymbol{\beta}_j = \gamma + \boldsymbol{u}_j$$

with regression coefficients $[\boldsymbol{\tau}_{ij}, \boldsymbol{\beta}_j, \gamma]$ that correspond to Compound Polynomial (CP) growth design elements for a given instructional week, $\boldsymbol{a}'_{tij}$, and student covariates $\boldsymbol{X}_{ij}$ employed. The CP is designed to simultaneously estimate between- and within-year growth. Such an approach means that norms for student growth that account for seasonal patterns in achievement (e.g., summer loss) can be produced for gains comparing any between- or within-year timepoints.

While the CP is a fairly traditional growth model in terms of the parameters, the design matrix is quite unique. The CP, and in particular the design matrix, is described in greater detail in the Appendix.

When examining gaps, $\boldsymbol{X}_{ij} = blockdiag([1\ male_{ij}])$, with an entry for each student growth component, $\boldsymbol{\tau}_{ij}$. By comparison, when only looking at growth unconditional on biological sex, $\boldsymbol{X}_{ij} = \boldsymbol{I}_{ij}$, the equation follows that used in Thum and Hauser (2015). The residual errors e_{tij}, the student random effects $\boldsymbol{r}_{ij}$, and the school random effects $\boldsymbol{u}_j$ are assumed to be variates from multivariate normal distributions.

Using the selection matrices $\boldsymbol{H}_\pi$ and $\boldsymbol{H}_\beta$ to identify, respectively, the coefficients among $\boldsymbol{\tau}_{ij}$ and $\boldsymbol{\beta}_j$ which are random, leads to the mixed-effects formulation of the multivariate normal model for student i in school j

$$Y_{ij} \sim MVN(\boldsymbol{A}_{ij}\boldsymbol{X}_{ij}\gamma, \boldsymbol{\Sigma}_{ij}) \quad (8)$$

where the components of the variance for each level are

$$e_{tij} \sim N(0, \sigma^2_{tij}) \quad (9)$$

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

$$\boldsymbol{H}_\beta \boldsymbol{u}_j \sim MVN(0, \boldsymbol{T}_\beta)$$

and

$$\boldsymbol{\Sigma}_{ij} = \overbrace{\boldsymbol{A}_{ij}(\boldsymbol{X}_{ij}\boldsymbol{H}'_\beta)\boldsymbol{T}_\beta(\boldsymbol{H}_\beta\boldsymbol{X}'_{ij})\boldsymbol{A}'_{ij_i}}^{\boldsymbol{\Sigma}_\beta} + \overbrace{(\boldsymbol{A}_{ij}\boldsymbol{H}'_\pi)\boldsymbol{T}_\pi\left(\boldsymbol{H}_\pi\boldsymbol{A}'_{ij_i}\right)}^{\boldsymbol{\Sigma}_\pi} + \overbrace{diag(\boldsymbol{\sigma}^2_{ij})}^{\boldsymbol{\Sigma}_y} . \quad (10)$$

Developing Model-based Norms. Norms are defined by the predicted marginal and conditional multivariate normal distributions based on the estimates of fixed-effects $[\hat{\gamma}, var(\hat{\gamma})]$ and variance-covariance components $[\hat{\sigma}^2, \widehat{\boldsymbol{T}}_\pi, \widehat{\boldsymbol{T}}_\beta]$ from fitting the model given in the prior

equations. For example, achievement and growth student norms for males or females for any grade and term, or for any combination of grades and terms, may be obtained from

$$\text{MVN}\left(\mathbf{C}'\hat{\mathbf{Y}}, \mathbf{C}'\mathbf{A}\mathbf{X}\,\text{Var}\left(\hat{\boldsymbol{\gamma}}\right)\mathbf{X}'\mathbf{A}'\mathbf{C} + \mathbf{C}'\left[\hat{\boldsymbol{\Sigma}}_{\beta} + \hat{\boldsymbol{\Sigma}}_{\pi} + \hat{\boldsymbol{\Sigma}}_{y}\right]\mathbf{C}\right) \quad (11)$$

Where $\mathbf{C}'$ is a contrast matrix designed to select the time periods being compared, $\mathbf{A}$ is a polynomial growth design matrix, and $\mathbf{X}$ is a dummy variable equal to one for males. Similarly, achievement and growth school norms are given by

$$\text{MVN}\left(\mathbf{C}'\hat{\mathbf{Y}}, \mathbf{C}'\mathbf{A}\mathbf{X}\,\text{Var}\left(\hat{\boldsymbol{\gamma}}\right)\mathbf{X}'\mathbf{A}'\mathbf{C} + \mathbf{C}'\hat{\boldsymbol{\Sigma}}_{\beta}\mathbf{C}\right). \quad (12)$$

For additional details, please see Thum and Hauser (2015).

Finally, the ICCs for any grade, term, and value of the "male" dummy variable are easily obtained. Specifically (and following from Equation 12):

$$\mathbf{c}'\hat{\boldsymbol{\Sigma}}_{\beta}\mathbf{c} \Big/ \mathbf{c}'\left[\hat{\boldsymbol{\Sigma}}_{\beta} + \hat{\boldsymbol{\Sigma}}_{\pi}\right]\mathbf{c}. \quad (13)$$

Question 1. How different are effect size estimates of growth in student achievement when standardized relative to a distribution of gains rather than a distribution of scores from one timepoint? We begin by providing mean achievement, mean gains in achievement, and the SDs for each using our national sample (weighted to be nationally representative). As previously mentioned, since the longitudinal vertically scaled data we have are rare, we include these national norms for student achievement and growth based on those data. Such estimates can be used as empirical benchmarks for, say, program evaluations interested in how much treatment students grew relative to national averages for growth in the absence of intervention (just as we illustrate in Equation 3). Thus, even if readers prefer to use different effect sizes than the ones we present as empirical benchmarks, we give them the national means and variances of student test scores and score gains to produce the effect size of their choosing.

We then compare effect sizes for this question by producing them for inferences about samples and populations (empirical benchmarks). For the former, we used a generic test score gain of 5 RIT. One could think of this hypothetical as involving students in a particular grade, treatment group, or classroom who gained 5 RIT, on average. We then standardized those 5 RIT gains using Equations 1-3. As pointed out in Table 2, Equation 1 ignores the within-person correlation between timepoints (similar to the effect size produced by Bloom et al., 2008), Equation 2 uses the pooled SDs from the pre- and post-test but accounts for those correlations, and Equation 3 subtracts off "typical" growth before standardizing relative to the SD of the gains. Effect sizes are then compared. While such an approach is akin to a sensitivity analysis, as pointed out in Table 2, each effect size answers a slightly different research question. Therefore, we would not expect them to be equivalent. For population-level inferences, we looked at the gain for our national sample over the course of a year (pre- and post-test), then calculated an effect size using Equations 6-7.

Question 2. How different are effect size estimates of school-level growth when standardized relative to distributions of student- versus school-level gains?

We approached this question exactly as in Question 1, but using school level means[1] and SDs for achievement and growth. Thus, we provide these national school-level means and SDs for others who might wish to use them to produce an effect size. Also similar to Question 1, we produce effect sizes when the desired level of inference is a sample (again using a generic school-level gain of 5 RIT) versus the population. We then compare those effect sizes.

However, unlike in Question 1, the crux of this question involves comparing the effect sizes produced in Question 2 to those produced in Question 1. Given the means we report are from a three-level model and therefore represent school-level means, the main difference in the

effect sizes for Questions 1 and 2 is whether the denominator uses SDs for student versus school gains. Therefore, comparing results for the two questions means we can explore how consequential the decision is (commonly made in the literature per Bloom et al., 2008) to standardize school gains relative to distributions of student-level test scores rather than school-level SDs, which better match the estimand.

Question 3. What are national benchmarks for growth in male-female achievement gaps over time, and how sensitive are those empirical benchmarks to the variance used in the denominator?

For this question, we started by presenting means, mean gains, and the SDs of each for males, females, and all students together. Again, our sample is national and, consequently, such means and SDs can be used as empirical benchmarks if so desired. For example, states could compare changes in gaps in their schools to the national averages we provide. We then produce effect sizes for male-female gaps in test score gains using Equations 5 and 6, where the former does not account for within-person correlations and the latter does. Those effect sizes are compared. As previously noted, most studies examining achievement gaps as effect sizes use Equation 5, which will typically involve smaller SDs than Equation 6. Therefore, this question directly considers whether much of the gap literature may be understating changes in achievement gaps over time.

Results

Question 1. How different are effect size estimates of growth in student achievement when standardized relative to a distribution of gains rather than a distribution of scores from one timepoint?

Table 4 shows means and SDs of achievement and gains by grade in math, and Table 5 shows the same results in reading. One should note that horizontally (along the rows), the data are longitudinal. For example, the gain in 3rd grade, which is the gain between 2nd and 3rd grade, is for a cohort of students. However, down the column, the data are cross-sectional: the students in 4th grade are not the same as those who were in 3rd. This general format (longitudinal rows, cross-sectional columns) holds for all the subsequent tables in our study.

Examining inferences at the sample level with a generic 5 RIT gain, effect sizes differ considerably dependent on how they are calculated. Equation 2, which does not ignore the correlation between timepoints, produces effect sizes that are roughly .2 to .35 SDs greater than those produced in Equation 1, which does ignore that correlation. Thus, effect sizes in Equation 2, which we argue are less justifiable given high correlations in student achievement over time, tend to understate the effect size of student-level gains for a given sample.

INSERT TABLE 4

INSERT TABLE 5

Meanwhile, the difference-in-difference approach in Equation 3 answers an altered research question, namely how much the sample gained relative to the mean population gain in units of the SD of the gain. That is, this effect size helps answer: how much did the sample grow relative to typical growth? When using this metric, the effect sizes change considerably. Whereas an effect size that only examines the gain in Equation 2 can look quite large, the sample gain relative to normative growth tells a different story.

We also produced effect sizes in the case that inferences about the population using our national sample are of interest as empirical benchmarks. For example, we show that, using Equation 2, the 3rd to 4th grade gain is about .93 SDs, and those gains decrease steadily in SDs as

students move through school. One should note that these effect sizes are roughly .2 SDs larger than those based on Equation 1 ignoring the correlation in test scores over time.

Question 2. How different are effect size estimates of school-level growth when standardized relative to distributions of student- versus school-level gains?

Tables 6 and 7 show the same effect sizes as in Tables 4 and 5, but using school-level gains and SDs. As before, effect sizes are larger when accounting for over-time test score correlations. They also look much different when estimated relative to the typical school level gain (Equation 3).

INSERT TABLE 6

INSERT TABLE 7

More importantly, when comparing the student-level results in Tables 4 and 5 to the school-level effect sizes in Tables 6 and 7, results are substantively different. As anticipated, effect sizes are much larger for schools because the SDs of achievement and gains in achievement are much smaller at the school level. For example, most effect sizes in math based on Equation 2 are twice as large when based on the school SD rather than the student SD. In some cases, there is a 1 SD difference in the effect sizes. Thus, effect sizes that standardize school-level estimates relative to student-level SDs are likely understating the magnitude of school-level growth considerably.

Question 3. What are national benchmarks for growth in male-female achievement gaps over time, and how sensitive are those empirical benchmarks to the variance used in the denominator?

Table 8 presents effect sizes for the achievement gap in mathematics gains between males and females. While females make the largest gain between 2nd and 3rd grade, males make

larger gains thereafter, suggesting the initial gap in math favoring boys in 3rd grade may widen over time (again, these patterns are not reflected down the columns, which are cross-sectional whereas the gains on the rows are longitudinal). The effect sizes using Equation 6, which accounts for the across-time correlation, range from .04 SDs in 3rd grade (favoring females) to -.07 SDs (favoring males). These effect sizes are often twice as large as those that use Equation 5, which does not account for the over-time correlation of test scores. Thus, effect sizes based on Equation 5 likely understate the magnitude of the gap.

INSERT TABLE 8

Table 9 presents the same findings in reading. Like in math, gains initially favor females, but favor males by 5th grade. Given females start with higher achievement in 3rd grade, such a pattern could suggest that males are closing the gap as they move through school. Also similar to math, effect sizes for the gains are roughly 1.5 to 2 times as large when using Equation 6 compared to Equation 5.

INSERT TABLE 9

Discussion, Limitations, and Next Steps

Estimating student achievement growth and schools' contribution to that growth is increasing in prevalence in education practice and policy. For example, under *The Every Student Succeeds Act* (ESSA) of 2015, many schools are being held accountable for contributions to student growth over time. Further, with consistent attention paid to achievement gaps and their development over time (Fryer & Levitt, 2006; Robinson & Lubienski, 2011; Soland, 2018), comparing rates of differential growth to meaningful empirical benchmarks is increasingly necessary (Bloom et al., 2008; Hill et al., 2008). Given that most related studies use test scores with different scales, putting results on a common metric that is comparable across sample sizes

is important. Typically, effect sizes that take some numerator and divide by some variance, as with some variant of Cohen's *d*, are employed for that purpose, though there are several other effect sizes we do not discuss (e.g., Hedges & Hedburg, 2007).

Yet, most effect sizes for growth are standardized using variances from test scores at a point in time or pooled SDs from two timepoints. By contrast, and as we show in this study, an alternative would be to standardize in a way that accounts for the within-person correlation in test scores over time (akin to a paired sample t-test), which are often high (~.70 in our sample). Yet another approach is to standardize by subtracting off "typical" growth in the numerator and dividing by a distribution of student- or school-level gains. This last approach arguably better aligns with the intended inference for intervention effects that are over and above, or net, the typical magnitude of gains that naturally occur in the background. Thus far, few studies consider the implications of these choices about how to standardize achievement gains, nor discuss the different ways of interpreting the various effect sizes. In so doing, we provide several relevant results.

First, we provide a taxonomy of effect sizes for achievement gains in Table 2, including providing interpretations of each effect size relative to a given research question. As the table shows, one should begin by determining whether the effect size (a) is for one group versus a comparison of two groups, (b) is for a group of students or schools, and (c) whether the desired level of inference is for a sample, or is meant to be used as an empirical benchmark that approximates population values. Once those determinations have been made, several effect sizes are available. In some cases, the choice of effect size is merely related to the research question of interest. For example, whereas one effect size for a single group of students/schools in a sample involves standardizing the gain relative to the pooled SDs in a way that accounts for the

over-time correlation in scores (Equation 2), another involves subtracting off the mean population gain representing typical growth and standardizing relative to a distribution of gains (Equation 3). Both are valid, and the latter helps answer the question: how much did, say, students who underwent an intervention grow relative to typical growth in SDs of the gain?

While these two effect sizes are both reasonable, we would argue not all effect sizes are equally justifiable. For example, effect sizes that do not incorporate the correlation in scores over time—which are quite high—are likely to understate the effect sizes for student gains, and are therefore less justifiable. Yet, the vast majority of studies on student learning over time ignore that correlation. As evidence of how ignoring over-time correlations can affect inferences, we show that effect sizes for student gains are typically much smaller when the denominator is the pooled SD of test scores from two timepoints rather than the SD of the gains between those two points. For example, using a generic gain of 5 RIT as a comparison point, effect size for growth in math are often .2-.3 SDs larger when using an effect size that accounts for the correlation in scores over time. A similar phenomenon occurs when comparing estimates of population level growth by grade level for potential use in empirical benchmarks. Therefore, one could argue that effect sizes for student growth typically shown in the literature, including for use as empirical benchmarks (e.g., Bloom et al., 2008), understate growth that occurs each year because they do not account for the oftentimes high association between within-student test scores over time.

Beyond over-time correlations, our analyses point to another issue in much of the effect size literature, namely that there is natural level of growth that one might wish to net out when evaluating the impact of an intervention. That is, we produce effect sizes (Equations 3-4) designed to address the research question: How large is a particular gain over and above the

growth that naturally occurs between time points? This question is especially relevant to program evaluation. Typically, effect sizes for a gain are standardized relative to a distribution of test scores. Yet, another approach is to subtract off the typical population gain before standardizing. Thus, one can determine not just how large a gain was for students participating in a program, but how large that gain was relative to normal growth. We would argue that, in many program evaluation contexts, subtracting off the typical gain is closer to the estimand of interest. Our results show that such effect sizes lead to very different inferences compared to those that do not subtract of typical gains.

We also show that, leaving aside the particular effect size, results will be very different when the numerator is school-level gains and the denominator is based on a student-level distribution rather than a school-level distribution. While this point has been made before theoretically (e.g., Bloom et al., 2008) and may be obvious mathematically, the degree to which such a mismatch between the numerator and denominator has not been examined, especially using a national dataset. In our analyses, effect sizes for school-level growth are twice as large when the denominator is a distribution of school-level rather than student-level gains. This result indicates that the fairly common practice of standardizing school gains relative to the variance in student-level test scores is likely to understate the magnitude of the impact a program or practice has had on that school.

These approaches to estimating effect sizes have additional implications for gaps in growth. We estimated gaps in male-female achievement growth, and showed that they are quite different depending on whether the denominator does or does not account for an over-time correlation. In practical terms, effect sizes that use pooled SDs from two timepoints likely understate the magnitude of the male-female gap in growth. Beyond the limitations of effect

sizes that do not account for correlated scores, one could argue that effect sizes with the SD of the gains in the denominator are more relevant to most gaps-based policy questions of interest. For instance, one could argue that the desired estimand is how much females have gained in math relative to males in SDs of the typical overall gain, not how much males are gaining relative to females in SDs of the baseline test score. That is, we want to know: how much are females gaining relative to typical growth for that grade, or relative to how much males have gained? Using the SD of the gains in the denominator gets us closer to that question.

Given our achievement gap results use a national sample weighted to be nationally representative, one can also use our results as empirical benchmarks for evaluating changes in male-female gaps as students progress through school (Bloom et al., 2008; Hill et al., 2008; Lipsey et al., 2012). That is, we provide policy-relevant empirical benchmarks for male-female gaps that can be used to determine practical significance in the ways described in prior effect size literature (Bloom et al., 2008; Hill et al., 2008; Lipsey et al., 2012). However, unlike that prior literature, we use longitudinal data to present means and SDs for not only achievement, but gains in that achievement. Thus, our results can be useful to researchers and evaluators trying to understand whether the changes in achievement gaps they see in a particular context are large or small relative to typical changes in those gaps nationwide.

Broader Implications for Reporting Effect Sizes and Limitations

Given our findings, what should be done? How should effect sizes, especially those for growth, be reported going forward? On one hand, our theoretical framework and results collectively help make the case that standardizing growth relative to a distribution of growth is preferable to standardizing with SDs from one timepoint as the variance. Doing the latter ignores the covariances between within-person test scores over time, which are often large (~.70

in our national sample) (Gibbons et al., 1993; Zimmerman, 1997). Thus, effect sizes for growth are likely understated. Further, effect sizes that use growth in the numerator and denominator are arguably closer to the estimand of interest. In the case of a program evaluation, we would argue that the real question of interest is how much a group of students has grown in the presence of an intervention compared to typical growth for all students. Using the variance in student gains in nationwide as the denominator is exactly what the norms we present do.

On the other, there is at least one compelling counterargument. Specifically, while the benchmarks created by Cohen are at best a mismatch and at worst misleading when used inappropriately, they are quite common and constitute a simple rule of thumb. (The same could be said for using pooled SDs in the denominator, which are common.) By contrast, using our effect sizes for growth can produce quite large values. For example, some analyses we did (not reported in the study) when examining school-level growth between years produced effect sizes of 2 SDs or higher on a regular basis. This phenomenon likely occurred because gains for schools over time can be fairly modest relative to the distribution of overall test scores at a given timepoint. Our approach would mean internalizing a new metric that might take some time to become familiar. If the purpose of the entire endeavor is to increase the interpretability of effect sizes, then such considerations are worth pondering.

An additional limitation is that being able to generate the effect sizes we do is dependent on having an equal-interval cross-grade scale as we do in MAP Growth. (One should note that in many of the studies we reviewed, an interval scale is assumed without much justification or explanation.) Oftentimes, program evaluators and policy analysts interested in questions like gaps are not so lucky, which limits the applicability of our study in some regards. Even if one does have a test that arguably possesses those properties, debates over what "equal interval"

means are still ongoing, and often heated (Ballou, 2009; Ho, 2009; Soland, 2017). Thus, producing effect sizes like ours may not be practicable or desirable in some cases.

Ultimately, the most sensible approach is likely to report all three effect sizes (Equations 1-3 for a sample of students or schools) when scores from an appropriate measure are available to standardize relative to a distribution of gains. Such a decision would mean an effect size that falls within a more commonly accepted range (e.g., <1) is available (Equation 1). However, evaluators and policy analysts could also get a more appropriate effect size that uses growth in the denominator. Further, our results (and those from the norms produced by Thum and Hauser, 2015) can be used as empirical benchmarks against which to compare those effect sizes that standardize relative to growth can be compared. The best way to make the effect sizes we suggest more easily understood is to provide nationally weighted averages for those effect sizes so that researchers have an empirical benchmark in the event that they do not have an internalized sense of how large an effect size needs to be in order to be considered practically significant. The norms developed by Thum and Hauser (2015) were created in part for that purpose.

Future Directions

Given the complexities associated with estimating student growth, there are many possible extensions to this work. To us, four stand out. First and most basically, effect sizes in the Cohen's *d* family are really just one type of effect size. Measures like correlations and odds ratios are also forms of effect sizes (see, e.g., Rosenthal & Rosnow, 1981). These different effect sizes are often associated with coarse empirical benchmarks of their own, such as arbitrary cutoffs for how large a correlation needs to be to achieve practical significance. More thought should be given to such effect sizes in a growth context.

Second, while our estimates of growth account for days of exposure to instruction, this issue is not a primary emphasis in this study. However, effect sizes and empirical benchmarks could be produced to show how much gains for students and schools change normatively dependent on days of instruction received throughout the year. For example, effect sizes for an intervention could be compared to normative gains for students experiencing the same days of educational exposure. Issues of how time is accounted for, coded, and conceptualized in the growth effect size literature merits attention.

Third, our results can generally be thought of as the marginal growth percentile for a given student or school. However, comparing growth is often not as meaningful for students or schools with marked differences in achievement at the pre-test. Thus, the norms developed by Thum and Hauser (2015) upon which analyses are based also include growth norms conditional on prior achievement. The appropriate effect sizes for conditional estimates of growth should be further explored.

Finally, additional empirical benchmarks should be produced for other policy metrics that have a growth component. For example, few strong empirical benchmarks exist for summer learning loss, including how patterns of achievement decline during the summer (summer loss) shift as students move through school. Similar benchmarks could be provided for Black-White and other race- and socioeconomic-based gaps.

Conclusion

In many regards, effect size and standardization have become synonymous. Mean differences are put on a common scale—usually in SDs—to ease comparisons across samples, sample sizes, and scales of measurement. In this study, evidence from our national longitudinal sample and accompanying norms model indicates that effect sizes for student growth can, in

certain circumstances, be made more germane to the practice or policy estimand of interest by paying attention to the variance used in the denominator of the effect size, and by subtracting of "typical" growth in the numerator when appropriate. That is, we try to adhere to the criterion articulated by Kelley and Preacher (2012) that an effect size should directly address a research or policy question of interest. Our results show that effect sizes for student and school growth are sensitive to whether they are standardized relative to static achievement versus growth, the latter of which accounts for a correlation between test scores at the two timepoints used. More importantly, we suggest that an effect size that not only accounts for that correlation, but also compares growth in a sample to normative growth overall is more policy relevant, especially in a program evaluation context. When evaluating a program, we would argue the estimand of interest is often how the growth of treated students or schools compares to typical growth. The growth effect sizes we produce better match this estimand by putting effect sizes in SDs of the distribution of gains, not static achievement.

Notes

1. One should note that school here is not defined the way people might typically define it. Rather, a school is a grade cohort of students in the same school, but it is not the full school.

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education*, *4*(4), 351–383.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328.

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).

Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review*, *8*(2), 249–281.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics*, *18*(3), 271–279.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201–228.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review*, *21*(1), 1–17.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., … Busick, M. (2012). *Translating the Statistical Representation of the Effects of Education Interventions Into*

More Readily Interpretable Forms. Washington, D.C.: Institute of Education Sciences, US Department of Education.

Quinn, D. M., Cooc, N., McIntyre, J., & Gomez, C. J. (2016). Seasonal dynamics of academic achievement inequality by socioeconomic status and race/ethnicity: Updating and extending past research with new national data. *Educational Researcher*, *45*(8), 443–453.

Robinson, J. P., & Lubienski, S. T. (2011). The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School Examining Direct Cognitive Assessments and Teacher Ratings. *American Educational Research Journal*, *48*(2), 268–302.

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.

Soland, J. (2018). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*, *120*(12).

Thum, Y. M., & Carl Hauser. (2015). *NWEA 2015 MAP Norms for Student and School Achievement Status and Growth.* Portland, OR: NWEA.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in Reading and Math Skills Forms Mainly before Kindergarten: A Replication, and Partial Correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357.

Yen, W. M. (2007). Vertical scaling and no child left behind. In *Linking and aligning scores and scales* (pp. 273–283).

Zimmerman, D. W. (1997). Teacher's corner: A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics*, *22*(3), 349–360.

Table 1(a)
Comparison of Cohort Versus Cross-sectional Measures of Mean Change

Grade	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
3	Cross-section 1/ Cohort 1					
4	Cross-section 1	Cohort 1				
5	Cross-section 1		Cohort 1			
6	Cross-section 1			Cohort 1		
7	Cross-section 1				Cohort 1	
8	Cross-section 1					Cohort 1

Table 2

Taxonomy of Effect Sizes for Growth in Achievement

One Group of Students/Schools	Desired Inference Level	Research Question How large is the gain in the sample/population relative to:
Eq. 1	Sample or Pop.	Pooled SDs of pre- and post-test scores?
Eq. 2	Sample or Pop.	Pooled SDs of pre- and post-test scores accounting for a pre/post correlation?
Eq. 3	Sample	The mean population gain from the same school year in SDs of the gain?
Eq. 4	Pop.	The prior school year population gain in SDs of the gain in the prior year?
Two Groups of Students/Schools	Desired Inference Level	How large is the difference in the sample/population gains relative to:
Eq. 4	Sample or Pop.	SDs of the pooled scores (both groups) from the pre-test?
Eq. 5	Sample or Pop.	SDs of the pooled gains (both groups) between the pre- and post-test?

Table 3

Mean Achievement, SDs by Grade in our Sample

		Mathematics			Reading		
Grade	Statistic	Fall	Winter	Spring	Fall	Winter	Spring
3	Mean	190.4	198.21	203.4	188.29	195.6	198.62
	School SD	5.88	6.12	6.46	6.7	6.59	6.63
	Student SD	13.1	13.29	13.81	15.85	15.14	15.1
	Ns	1456, 135458, 609075			1457, 134519, 606602		
4	Mean	201.94	208.71	213.49	198.16	203.6	205.92
	School SD	6.11	6.55	7.03	6.49	6.43	6.48
	Student SD	13.76	14.27	14.97	15.53	14.96	14.92
	Ns	1443, 130077, 592305			1456, 134361, 615399		
5	Mean	211.44	217.23	221.36	205.68	209.83	211.79
	School SD	7	7.57	8.16	6.27	6.23	6.31
	Student SD	14.68	15.33	16.18	15.13	14.65	14.72
	Ns	1440, 148818, 518378			1448, 148564, 526716		
6	Mean	217.62	222.06	225.32	210.99	214.21	215.75
	School SD	7.19	7.61	8.05	6.31	6.31	6.44
	Student SD	15.53	16	16.71	14.94	14.53	14.66
	Ns	1451, 165541, 546568			1452, 162887, 554445		
7	Mean	222.65	226.12	228.59	214.45	216.91	218.16
	School SD	7.71	8.06	8.43	6.57	6.57	6.69
	Student SD	16.59	17.07	17.72	15.31	14.98	15.14
	Ns	1415, 190705, 781050			1418, 194033, 814818		
8	Mean	226.3	229.15	230.93	217.24	219.09	220.07
	School SD	8.8	9.21	9.62	7.47	7.39	7.48
	Student SD	17.85	18.31	19.11	15.72	15.37	15.73
	Ns	1377, 199759, 619604			1396, 206667, 664327		

Table 4

Comparison of Different Effect Sizes for Mean Student Gains in Math

Grade	Point in Time		Gain		Sample Inference, 5 RIT Gain			Population Inference		
	Mean	SD	Mean	SD	Gain/Pooled SD	Gain/(Pooled SD - 2*cov)	(Gain - Mean Gain)/SD of the Gain	Gain/Pooled SD	Gain/ (Pooled SD - cov)	(Mean Gain T2 - Mean Gain T1)/SD of the Gain
					Eq. 1	Eq. 2	Eq. 3	Eq. 1	Eq. 2	Eq. 4
3	191.59	13.62	11.81	6.76	X	X	X	X	X	X
4	203.54	14.12	9.95	6.63	0.36	0.57	-0.75	0.72	1.14	-0.28
5	212.74	15.77	8.62	7.01	0.33	0.70	-0.52	0.58	1.20	-0.20
6	220.44	17.35	4.88	7.15	0.30	0.65	0.02	0.29	0.63	-0.53
7	223.68	17.78	4.92	6.59	0.28	0.56	0.01	0.28	0.55	0.01
8	226.92	18.55	4.00	7.77	0.28	0.52	0.13	0.22	0.42	-0.14

Table 5

Comparison of Different Effect Sizes for Mean Student Gains in Reading

Grade	Point in Time		Gain		Sample Inference, 10 RIT Gain			Population Inference		
	Mean	SD	Mean	SD	Gain/Pooled SD	Gain/(Pooled SD - 2*cov)	(Gain - Mean Gain)/SD of the Gain	Gain/Pooled SD	Gain/ (Pooled SD - 2*cov)	(Mean Gain T2 - Mean Gain T1)/SD of the Gain
					Eq. 1	Eq. 2	Eq. 3	Eq. 1	Eq. 2	Eq. 4
3	188.15	16.25	10.47	7.25	X	X	X	X	X	X
4	198.39	15.97	7.54	6.83	0.31	0.51	-0.37	0.47	0.77	-0.40
5	205.83	15.57	5.96	7.19	0.32	0.55	-0.13	0.38	0.65	-0.23
6	211.55	15.92	4.20	7.48	0.32	0.59	0.11	0.27	0.50	-0.24
7	214.33	16.19	3.83	7.05	0.31	0.56	0.17	0.24	0.43	-0.05
8	216.99	15.89	3.09	8.30	0.31	0.57	0.23	0.19	0.35	-0.10

Table 6

School Level Comparison of Different Effect Sizes for Mean Student Gains in Math

Grade	Point in Time		Gain		Sample Inference, 5 RIT Gain			Population Inference		
	Mean	SD	Mean	SD	Gain/Pooled SD	Gain/(Pooled SD - cov)	(Gain - Mean Gain)/ SD of the Gain	Gain/Pooled SD	Gain/ (Pooled SD - cov)	(Mean Gain T2 - Mean Gain T1)/SD of the Gain
					Eq. 1	Eq. 2	Eq. 3	Eq. 1	Eq. 2	Eq. 4
3	191.59	6.29	11.81	2.74	X	X	X	X	X	X
4	203.54	6.48	9.95	2.44	0.78	1.24	-2.03	1.56	2.47	-0.68
5	212.74	7.78	8.62	2.85	0.70	1.40	-1.27	1.20	2.42	-0.55
6	220.44	8.11	4.88	2.53	0.63	1.38	0.05	0.61	1.34	-1.31
7	223.68	8.50	4.92	2.20	0.60	1.18	0.04	0.59	1.16	0.02
8	226.92	9.18	4.00	2.50	0.57	1.06	0.40	0.45	0.85	-0.42

Table 7

School Level Comparison of Different Effect Sizes for Mean Student Gains in Reading

Grade	Point in Time		Gain		Sample Inference, 5 RIT Gain			Population Inference		
	Mean	SD	Mean	SD	Gain/Pooled SD	Gain/(Pooled SD - cov)	(Gain - Mean Gain)/ SD of the Gain	Gain/Pooled SD	Gain/ (Pooled SD - cov)	(Mean Gain T2 - Mean Gain T1)/SD of the Gain
					Eq. 1	Eq. 2	Eq. 3	Eq. 1	Eq. 2	Eq. 4
3	188.15	6.85	10.47	1.95	X	X	X	X	X	X
4	198.39	6.80	7.54	1.74	0.73	1.16	-1.46	1.10	1.75	-1.50
5	205.83	6.63	5.96	1.80	0.74	1.59	-0.53	0.89	1.90	-0.91
6	211.55	6.80	4.20	2.20	0.74	1.63	0.36	0.63	1.37	-0.98
7	214.33	7.16	3.83	2.09	0.72	1.41	0.56	0.55	1.08	-0.17
8	216.99	7.36	3.09	2.24	0.69	1.31	0.85	0.43	0.81	-0.35

Table 8

Male-female Empirical Benchmarks: Mean Achievement, Gains, and Effect Sizes in Math

Grade	Male				Female				Overall				Male - Female	
	Mean	S.D.	Mean Gain	S.D. Gain	Mean	S.D.	Mean Gain	S.D. Gain	Mean	S.D.	Mean Gain	S.D. Gain	Mean Diff./Pooled SD	Mean Diff./SD of Gains
													Eq. 5	Eq. 6
3	203.70	14.00	11.92	6.78	203.09	13.60	11.69	6.46	203.40	13.81	11.81	6.76	0.02	0.04
4	213.65	15.16	9.83	6.63	213.32	14.77	10.08	6.62	213.49	14.97	9.95	6.63	-0.02	-0.04
5	221.37	16.38	8.37	7.02	221.35	15.97	8.88	6.99	221.36	16.18	8.62	7.01	-0.03	-0.07
6	225.26	16.92	4.65	7.18	225.40	16.48	5.12	7.12	225.32	16.71	4.88	7.15	-0.03	-0.06
7	228.39	17.88	4.71	6.60	228.83	17.54	5.14	6.58	228.58	17.72	4.92	6.59	-0.02	-0.07
8	230.70	19.26	3.96	7.68	230.41	18.93	4.19	7.65	230.93	19.11	4.00	7.77	-0.01	-0.03

Table 9

Male-female Empirical Benchmarks: Mean Achievement, Gains, and Effect Sizes in Reading

Grade	Male				Female				Overall				Male - Female	
	Mean	S.D.	Mean Gain	S.D. Gain	Mean	S.D.	Mean Gain	S.D. Gain	Mean	S.D.	Mean Gain	S.D. Gain	Mean Diff./Pooled SD	Mean Diff./SD of Gains
													Eq. 5	Eq. 6
3	197.34	15.14	10.63	7.01	199.97	14.80	10.31	6.95	188.15	16.25	10.47	7.25	0.02	0.05
4	204.62	14.99	7.63	6.57	207.28	14.61	7.45	6.52	198.39	15.97	7.54	6.83	0.01	0.03
5	210.59	14.77	5.90	6.94	213.06	14.44	6.02	6.90	205.83	15.57	5.96	7.19	-0.01	-0.02
6	214.33	14.68	4.03	7.27	217.25	14.35	4.40	7.17	211.55	15.92	4.20	7.48	-0.03	-0.05
7	216.63	15.14	3.66	6.83	219.80	14.82	3.99	6.72	214.33	16.19	3.83	7.05	-0.02	-0.05
8	218.45	15.73	3.03	8.06	221.86	15.41	3.30	7.97	216.99	15.89	3.09	8.30	-0.02	-0.03

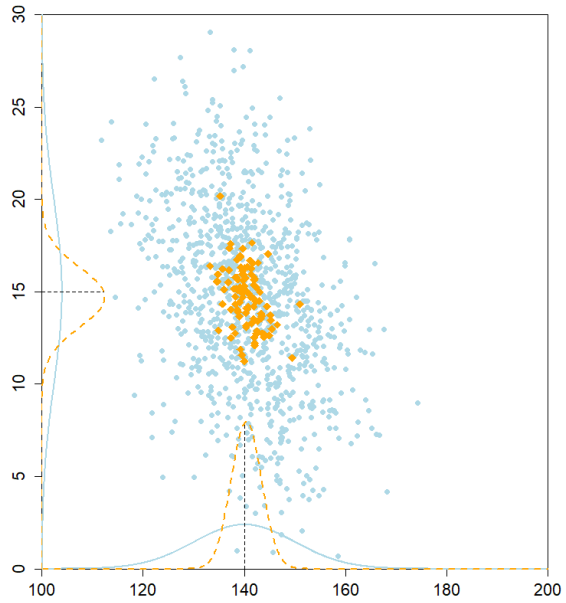

Figure 1. Units are RIT points, with mean 3rd grade RIT on the horizontal axis and mean 3rd to 4th grade gain on the vertical axis. Student scores/gain scores are represented by blue dots and school-level mean scores/gain scores by orange dots. Distributions for mean achievement and gains for students versus schools (also color-coded) are superimposed on the axes.

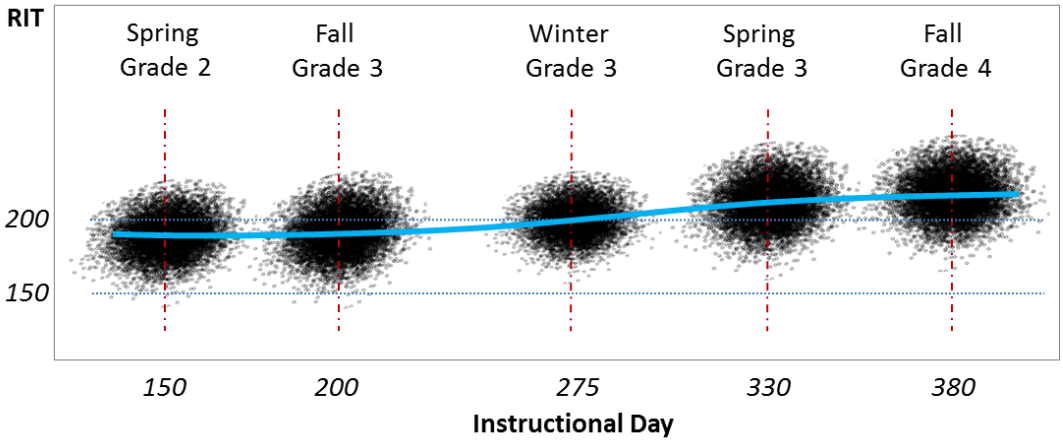

Figure 2. A schematic diagram of student longitudinal data employed in growth models for students beginning in grade three.

Appendix A: Post-stratification Weights

Data in our sample are from a large archive of longitudinal student assessment results. However, its contents constitute only a convenience sample. To better approximate norming results that reflect the U.S. school-age student population, we use the School Challenge Index (SCI), an indicator of how schools in a state varied in terms of the challenges and opportunities they operate under as reflected by an array of school-level factors they do not control. The SCI is a broad proxy for student poverty at the school. The index ranges from 1 to 99, and higher SCI schools tend to be those with a greater concentration of disadvantaged students as a proportion of their student body.

Broadly, the SCI is designed to measure the level of challenge a school faces in terms of the socio-economic composition of its student body and as moderated by other related factors such as the ethnic make-up of its students, the kind of assistance they receive (e.g., Title 1), or the environment it operates within such as its location (e.g., city, rural), size, school level (e.g., elementary, high school), and emphasis (e.g., magnet, charter). Tables A1 and A2 present descriptive statistics for each variable used in the SCI for MAP and non-MAP users. These data come from the 2008 National Center for Educational Statistics (NCES) Common Core of Data in the Public School Universe (NCES-CCD-PSU) dataset (Sable & Plotts, 2010).

To construct school-level weights from the SCI, we first identify schools by their SCI deciles. If P_d^{NWEA} is the proportion of NWEA schools and P_d^{ALL} is the proportion of all schools in SCI decile d, the population weight for school j in SCI decile d is

$$W_j = \frac{P_d^{ALL}}{P_d^{NWEA}}$$

This weighting procedure is repeated for each subject and grade level because not all MAP users test on the same subject or serve the same grade levels. These weights are then employed as

post-stratification weights in the growth model. For additional details on the construction of the SCI and weighting procedures, please see Thum and Hauser (2015).

Appendix B: Building a Compound Polynomial Design Matrix

Building the CP design matrix begins by specifying within- and between-year design matrices. In this illustration, we will assume five years of data are available. 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

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

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

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

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 first year, and the values in the first two columns and the second row represent spring of the first year. Similarly, the values in the last two columns in the last row represent spring of the fifth year.

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, $\boldsymbol{D_b}$, 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 $\boldsymbol{D_b}$ =

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

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.

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

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

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**:

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

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

In this matrix, the first three columns represent the intercept, linear growth, and quadratic growth terms for the between-year spring-to-spring component. These are like the components of a traditional polynomial growth model with only spring data. In contrast, 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 B1 below. This design matrix can then be used in the growth model described in the methods section.

In that growth model, the coefficients on columns 4-6 in the **CP** matrix capture within-year growth. The intercept (column 4) is the predicted fall-to-spring growth in the centering year, the coefficient on Column 5 is the linear growth rate of change for fall to spring growth, and the coefficient corresponding to Column 6 is the quadratic term for that growth. That is, the first coefficient (intercept) 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.

Table A1.
Distributions of Categorical NCES Variables Contributing to the School Challenge Index by NWEA School Affiliation

		NWEA Affiliated?				
		Yes		No		
SCI Variable	Category	N	%	N	%	Total
School locale	City	5,321	23.7%	18,945	27.0%	24,266
	Suburb	5,700	25.4%	20,070	28.6%	25,770
	Town	3,318	14.8%	8,992	12.8%	12,310
	Rural	8,108	36.1%	22,261	31.7%	30,369
School type	Regular	21,578	96.1%	65,514	93.2%	87,092
	Special Ed.	191	0.9%	1,337	1.9%	1,528
	Alternative/other	678	3.0%	3,417	4.9%	4,095
School level	Elementary	12,999	57.9%	39,336	56.0%	52,335
	Middle	4,243	18.9%	12,156	17.3%	16,399
	High	4,146	18.5%	14,334	20.4%	18,480
	Other	1,059	4.7%	4,442	6.3%	5,501
Title 1 eligible	Yes	16,343	72.8%	51,033	72.6%	67,376
	No	5,525	24.6%	16,438	23.4%	21,963
	Missing/not reported	579	2.6%	2,797	4.0%	3,376
Magnet school	Yes	702	3.1%	2,188	3.1%	2,890
	No	15,740	70.1%	52,207	74.3%	67,947
	Missing/not reported	6,005	26.8%	15,873	22.6%	21,878
Charter School	Yes	1,603	7.1%	3,843	5.5%	5,446
	No	16,940	75.5%	60,951	86.7%	77,891
	Missing/not reported	3,904	17.4%	5,474	7.8%	9,378

Table A2.
Distributions of Continuous NCES Variables Contributing to the School Challenge Index by NWEA School Affiliation

	NWEA Affiliated?						
	Yes			Yes			
SCI Variable	Mean	SD	N	Mean	SD	N	Total
Proportion free/reduced price lunch	0.48	0.29	22,347	0.50	0.95	69,426	91,773
Proportion racial minority	0.33	0.32	22,447	0.40	0.33	70,268	92,715
Staff full-time equivalents	3.15	0.79	22,447	3.18	0.93	70,268	92,715