# THE EFFECT OF SUMMER LEARNING LOSS ON ANNUALLY ESTIMATED STUDENT GROWTH PERCENTILES

Susan Lyons

January 2017



Lyons
ASSESSMENT
CONSULTING

## ABSTRACT

Despite widespread adoption of student growth percentiles (SGPs) to link student growth to educator and school evaluations, little empirical research exists addressing the validity of the estimates for use in high-stakes personnel evaluation systems. This is especially troublesome given the moderate correlations reported between mean student growth percentiles (MGPs) and select student characteristics, specifically poverty, which have been well documented. This study explores summer learning loss (SLL) as one potential source of bias in MGP estimates that could help explain the relationship between poverty and MGPs. The guiding research hypothesis is that unaccounted for variance in summer learning patterns are contributing to error variance in aggregate estimates of student growth. Data from two, widely-used interim assessment programs were analyzed and results reveal that while the effect of SLL on the validity of SGPs for evaluation purposes varies across the datasets, the correlations between SGPs and student poverty cannot be primarily explained by differences in summer learning patterns. Policy implications are discussed.



Suggested Citation: Lyons, S. (2017). *The Effect of Summer Learning Loss on Annually Estimated Student Growth Percentiles.* Lyons Assessment Consulting.

# TABLE OF CONTENTS



TABLE OF CONTENTS

# INTRODUCTION

Before the December 2015 passage of the *Every Student Succeeds Act* (ESSA), federal policy initiatives had been increasingly pushing for the use of student growth data in state-level teacher evaluation systems. Because states were not meeting the achievement target of 100% proficiency, as prescribed by the 2001 reauthorization of the *Elementary and Secondary Education Act*, the Obama administration implemented a flexibility waiver program. In exchange for relaxing the 100% proficiency mandate, flexibility waivers required that states implement a number of new accountability reforms including a strong emphasis on using student growth data for school personnel evaluation purposes. Specifically, the ESEA flexibility waivers required that student achievement data be used to discriminate effective teachers from less effective teachers in order to advise personnel decisions such as retention, promotion, and dismissal (U.S. Department of Education, 2012). Because full implementation of growth-based evaluation systems was expected for the 2016-2017 school year (U.S. Department of Education, 2013) states are currently in various stages of revising their teacher evaluation systems to include measures of student growth. With the recent passage of the *Every Student Succeeds Act*, states are now under no obligation to move forward with the educator evaluation systems laid out in their waiver agreements. Instead, states have a renewed opportunity to reconsider their assessment, evaluation, and accountability systems under a new, more flexible law. In redesigning these systems, states may continue to be interested in holding teachers and schools accountable for student growth, but may have concerns about the validity of such methods given a degree of skepticism in the field. This paper is intended to contribute to the literature on the validity of SGPs for use in educator and school evaluation systems. Additionally, the concluding section of this paper offers insight on the policy implications of the findings given the renewed flexibility offered under ESSA.

While a variety of methods have emerged to include measures of student growth in educator evaluations (e.g., value-added models, value tables, student learning objectives), the focus of the current study is the student growth percentile (SGP) model (Betebenner, 2008). The SGP model uses quantile regression to rank student achievement relative to peers with the same prior achievement scores. Student growth percentiles are aggregated at the teacher, grade and school levels using the mean or median of all student SGPs (at the aggregate point: i.e., teacher, grade, school, etc.) to yield a single summary score for the aggregated unit, an MGP. Research has shown that mean-based MGPs have more desirable statistical properties over median-based MGPs, mean-based MGPs are used in the current study and MGP is used henceforth to refer to the mean (Castellano & Ho, 2012). Unlike other growth models used for teacher evaluation, such as many value-added models (VAMs), SGPs condition only on prior achievement—often multiple years of prior achievement—and leave out other student- or school-level covariates such as demographic and socioeconomic information. Because of this feature, SGPs are generally considered to be a description of how much students have grown relative to peers with similar prior achievement; their intention is not to measure teacher effectiveness or isolate the portion of student growth that can be attributable to the teacher (Betebenner, 2008). Instead, MGPs derived from the SGP model are designed to be descriptive indicators for use within a more comprehensive evaluation system and *not* a direct measure of teacher quality. MGPs are designed to be considered within the context of the teaching environment and in relation to other indicators of teacher quality rather than as standalone quantitative measure to be used as a "total score" for teacher, grade, or school effectiveness. Though MGPs are operationalized

meaningfully differently than VAMs, the resultant estimates tend to correlate highly with coefficients greater than .8 (Goldhaber, Walch, & Gabele, 2014; Ehlert, Koedel, Parsons, & Podgursky, 2013).

Student growth percentiles have become a popular growth model for teacher evaluation and are now used as part of the accountability systems in at least 27 states (McEachin & Atteberry, 2014). In spite of their growing use, there is no lack of skepticism among educational researchers about the validity of student growth modeling for use in teacher evaluations (Braun, Chudowski, & Koenig, 2010; Lissitz, 2012; Sireci, 2013; Haertel, 2013). One of the most apparent validity concerns is the tendency of teacher and school MGPs to correlate with student characteristics. Goldhaber, Walch, and Gabele (2014) warn that differences in the make-up of students within the classroom can affect teacher-level MGPs considerably. These authors find that as the average prior achievement level of the classroom increases by one standard deviation, predicted MGPs increase by 35 percentile points in reading (e.g., predicted MGP changes from a 50 to a 85 based on student prior achievement alone) and 15 percentile points in math. Additionally, Wright (2010) finds statistically significant correlations from -0.13 to -0.31 between teacher-level MGPs and proportion of students eligible for free- or reduced-priced lunch (FRL) in the classroom. At the school level, Ehlert, Koedel, Parsons, and Podgursky (2013) find that disadvantaged schools are disproportionately underrepresented in the top quartile of schools on the median SGP metric. The underlying causes of these observed correlations, however, are unknown. The extent to which these correlations reflect an uneven distribution of teacher quality across students and schools, or are rather, manifestations of model bias is not yet clear and thus not well understood. If the SGP model, which conditions only on prior achievement, is insufficient for explaining all differences in student growth that are unrelated to teacher or school effectiveness, then the exclusion of other explanatory factors may be leading to omitted-variable bias in the MGP educator effectiveness estimates. In the next section, we present and discuss a rationale and model for understanding the possible causal influences for the observed correlations between MGPs and student-level characteristics, and thus a fundamental premise for this investigation.

## Defining the Problem

Gelman and Imbens (2013) posit that research in the social sciences is too focused on "effects of causes" rather than "causes of effects." Traditional behavioral research methodology often involves artificially creating or imposing a particular cause in order to measure the effect, rather than trying to understand the potential cause(s) of an observed effect. Integrating what Gelman (2011) calls "reverse causal inferences," or the search for causes, into the traditional research framework is one, often neglected, way to formalize scientific inquiry as a model-building process. In the context of the current study, the implicit model would be that aggregate SGPs are correlated with student level characteristics. Gelman and Imbens (2013) encourage researchers to question the implicit model and ask "why?" What could be causing the observed correlation? The purpose of posing this question is to improve the implicit model by identifying sources of confounding variables.

Because SGPs rank student achievement relative to their peers with comparable prior achievement, the effect of the modeling necessarily results in an orthogonal relationship between estimated student growth and previous observed achievement. In other words, no matter where students lie on the achievement continuum, they have equal probabilities of obtaining all possible growth scores. Therefore, it cannot be that the observed relationships at the aggregate levels are due to the presumably different absolute growth

trajectories at different points along the achievement scale. Given this, two alternative models are possible (as adapted from Gelman and Imbens, 2013):

$$Y_i(x) \perp Z_i, \text{ and} \quad (1)$$
$$Y_i \perp Z_i \mid V_i \quad (2)$$

where, in the context of the current study, $Y_i$ is an individual teacher, grade, or school MGP, $Z_i$ represents the student-level characteristics in a classroom/grade/school, and $\perp$ represents an orthogonal relationship. Let $Y_i(x)$ represent all possible outcomes or manifestations of the combination of variables $Y_i$ and $X_i$, each of which is denoted as $Y_i(X_i)$. In model 1, the observed association between the MGPs, $Y_i$, and student characteristics, $Z_i$, is an artifact of the causal effect of $X_i$ on $Y_i$ and a correlation between $X_i$ and $Z_i$. An example of variable $X_i$ may be the wealth of the school district. A school district that can afford to pay teachers high salaries and therefore is likely able to be more selective in its hiring of teachers may have, on average, higher-quality teachers with generally high MGP scores. Additionally, district wealth is not unrelated to student-level characteristics such as income and demographic variables. Therefore, the observed correlation between $Y_i(x)$ and $Z_i$ may be a function of district wealth, $X_i$. In other words, MGPs and student characteristics are by their nature unrelated, except for the reality that district wealth—itself a function of student characteristics—may be attracting teachers of higher quality as signaled by systematically higher MGPs. Alternatively, in model 2, the observed correlation is a result of the effect of a third variable, $V_i$, which has been omitted from the implicit model. In this case, $V_i$ would be a confounding variable that is related to both $Y_i$ and $Z_i$, and, when included, the relationship between $Y_i$ and $Z_i$ disappears or diminishes.

Using a causal inference framework (Gelman and Imbens, 2013), two, not mutually exclusive possible causes for the observed relationship between student-level characteristics, such as poverty, and MGPs are discussed: 1) an uneven distribution of teacher quality, and 2) an artifact of omitted-variable bias in the model. The first possible cause of the observed relationship between teacher- and school-level MGPs and student characteristics is if higher-achieving students have teachers that are on average more effective. Because students and teachers are not randomly assigned to schools, it is likely that, among other factors, perceptions of teacher quality may influence the placement of both students and teachers into schools. Research has shown that low-achieving students have a higher likelihood of being in a school with less skilled teachers (Lankford, Loeb, & Wyckoff, 2002). Urban schools struggle to attract teachers; this signals the pool of potential teaching candidates is proportionally smaller for urban districts, and some of those who end up in these classrooms are inevitably less qualified with respect to experience, education, and certification (Jacob, 2007). Betts, Zau, and Rice (2003) find that the schools with the highest test scores have teachers with, on average, two-and-a-half times as many years of experience as teachers in low-achieving schools. These teachers are also twice as likely to hold master's degrees. High-status schools tend to hire better qualified teachers who can provide students with more rigorous learning materials and educational experiences (Darling-Hammond, 1996; Ladson-Billings & Tate, 1995; Oakes & Lipton, 1993). On top of this, schools with low-achieving students have lower teacher retention rates, and some research suggests that retention rates are particularly low for teachers with better qualifications (Boyd, Lankford, Loeb, & Wyckoff, 2005; Guarino, Santibañez, & Daley, 2006; Hanushek, Kain, & Rivkin, 2004). Trouble attracting and retaining teachers at lower performing schools could be contributing to an overall decline in the average quality of the teacher workforce at these schools and therefore likely be leading to an uneven distribution of teacher quality across all school settings (Clotfelter, Ladd, Vigdor, & Diaz, 2004; Darling-Hammond, 1995).

While it is now generally accepted that student achievement is correlated with background factors of students, and the observed relationship is not due to bias in the measurement alone (Coleman et al., 1966), the same cannot be said for the correlations between teacher effectiveness estimates and student background variables. The underlying factors explaining the shared variance between MGPs and student characteristics are still being studied. Recent studies have found that these correlations can be at least partly explained by measurement error (McCaffrey, Castellano & Lockwood, 2015; Shang, VanIwaarden & Betebenner, 2015). The current study seeks to understand yet another possibility for explaining this relationship: omitted variable bias in the model. It could be that omitted, confounding variables related to both MGPs and student-level characteristics are creating systematic bias in the model. As Braun, Chudowsky, and Koenig (2010) explain, "Bias refers to the inaccuracy of an estimate that is due to a shortcoming or incompleteness in a statistical model itself" (p. 43). In this context, bias would occur if teachers or schools who have the same true effectiveness—that are equally effective at eliciting achievement growth from their students—receive different MGP estimates due to factors outside of their control (e.g., demographic characteristics) because these variables are unaccounted for in the model. If student-level prior achievement is insufficient for fully capturing the growth trajectory of all students, then exclusion of other explanatory variables would lead to omitted-variable bias. If subgroups of students, such as those living in poverty, have the same prior achievement as their peers, but for a variety of factors associated with poverty they do not have the same growth trajectory, then failing to include an indicator of poverty in the model would lead to bias in the estimator. The idea is that such factors influence both student achievement and growth. Excluding such relevant factors from the SGP model would fail to account for any interaction that may exist between student characteristics and rates of growth (see McCall, Hauser, Cronin, Kingsbury, & Houser, 2006).

From an evaluation system design point-of-view, it is vital to disentangle the possible causes for the observed correlation between MGPs and student characteristics. Failing to do so may unfairly penalize teachers for working with low-income students, or, when student characteristics are included in the model without theoretical and empirical support for their inclusion, over-compensate for the effects of poverty which may inadvertently hide systematic inequities in access to high-quality teachers.

## Purpose and Research Questions

The purpose of this research is to explore one possible cause for the observed relationship between MGPs and student characteristics in order to improve the current implicit model. A potential source of bias in teacher effectiveness estimates is summer learning loss. Summer learning loss refers to the well-studied phenomenon that students from low-income families tend to lose academic achievement over the summer, while students from wealthier homes tend to continue to gain in achievement during the summer months (Entwisle & Alexander, 1992). Although this pattern has been documented extensively by scholars (see Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; McCombs et al., 2011), many student growth estimates used for teacher evaluation continue to be calculated using a 12-month growth window, without any adjustment for the summer interval. Thus, these annual estimates include the summer months over which educators have little-to-no control. Therefore, teacher and school MGPs are essentially absorbing the positive or negative influences the summer vacation period has on their students' achievement. Entwisle and Alexander (1992) caution that "with differences between schools measured annually, and with schools in season only part of the year, there may be serious misspecification of 'school effects'" (p. 82). More recently, Haertel (2013) explicitly cites summer learning loss as a potential cause of bias in

teacher effectiveness estimates derived from value-added models: "On average, reading scores from the previous spring will *underestimate* the initial autumn proficiency of students in more advantaged classrooms and *overestimate* the initial autumn proficiency of those in less advantaged classrooms. Even if the two groups of students in fact make equal *fall*-to-spring gains, the measured prior *spring*-to-spring gains may differ" (p. 17).

The research hypothesis for this study is that conditioning on prior achievement alone, as is done in the SGP model, may not be sufficient for capturing the different, and likely economically-moderated growth rates that occur during the summer months. Because SGP estimates are typically calculated annually, summer learning loss is hypothesized to be correlated with not only student characteristics but also with spring-to-spring MGPs. Accounting for the summer months, might therefore improve the implicit model with the hypothesized model—Equation (2). To test this model, two research questions structure the two studies in this paper:

1. What proportion of classroom variance in summer learning patterns can be accounted for by poverty?
2. Does controlling for changes in achievement over the summer months reduce the magnitude of the relationship between MGPs and student-level poverty?

# LITERATURE REVIEW

## Student Growth Percentiles

The SGP model was originally adopted by Colorado as part of the Growth Model Pilot Program and was accepted by the U.S. Department of Education in 2009 (U.S. Department of Education, 2009). Just as traditional percentiles in educational assessment normatively describe the location of student scores within the context of a peer group, Student Growth Percentiles normatively describe student growth. Student Growth Percentiles are calculated by conditioning current achievement on measures of prior achievement. The conditional distribution is used to make a probability statement about a student's current score, relative to peers with similar prior achievement histories. In the SGP model, conditional probability densities are estimated with quantile regression. The conditional quantile functions used to estimate SGPs are specified using R (R Core Team, 2015) with the SGP package (Betebenner, VanIwaarden, Domingue & Shang, 2016). Student Growth Percentiles offer a good analytic framework for looking at growth in that they are content- and scale-neutral. This means that even if the content or the score scale of the assessments changes across measurement occasions, the SGP model can provide useful information with a consistent interpretation (i.e., describing how well a student performed relative to peers with similar prior achievement). This kind of norm-referenced interpretation is often preferable to other kinds of growth inferences that are scale-dependent (e.g., vertical scale score differences) which can be difficult to interpret. Mean Student Growth Percentiles are calculated by averaging the SGPs of all of the students in the unit (e.g., classroom, grade, school). MGPs provide a description of average individual student gains relative to peers with similar growth. Just like SGPs, MGPs range from 1-99 where higher scores indicate high average individual growth compared to students with similar prior achievement.

The current study uses the SGP model rather than other growth models used for teacher evaluation for two main reasons. First, most other growth models (e.g., value-added models) condition on student characteristics such as poverty in addition to prior achievement when estimating teacher effects. Because the SGP model does not do this, it is important to explore to what extent omitted variable bias may be present. Unsurprisingly, MGPs have been shown to have stronger correlations with student demographic and socioeconomic variables than other popular growth models for teacher evaluation because these variables are unaccounted for in the model (Ehlert, Koedel, Parsons, & Podgursky, 2014). Secondly, the SGP model continues to grow in use and with such popularity, studies that provide validity evidence for the model are surprisingly scarce. A search of the peer-reviewed literature archived in the ERIC database for "value-added model" returns 322 results, while the same database returns only 62 results for "student growth percentile." The present study is intended to be a contribution to the growth model literature.

## Summer Learning Loss

Summer learning loss is a broad term in educational research that refers to the achievement growth patterns of students over the summer months. Research shows that changes in achievement during the summer are moderated by subject matter, prior achievement, and poverty. A 1996 meta-analysis of 13 studies found that summer learning loss tends to be more dramatic in mathematics than in reading with respective effect sizes of $d$ = -.14 and $d$ = -.05 (Cooper et al., 1996). This could be because families more often spend more time reading at home than promoting or practicing mathematics skills (Harris & Sass, 2009). More recently, Helf, Konrad, and Algozzine (2008) found no evidence of summer setback for reading as they did in math, and instead reported summer gains in reading, especially for those students

on the lower end of the achievement scale. On the contrary, Burkam, Ready, Lee and LoGerfo (2004) found evidence of mathematics gains over the summer months, with gains being highest for higher-income students. However, this study relies on data from the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K), to calculate summer learning from kindergarten to first grade, which means these findings may not be generalizable across all grades.

In addition to being moderated by subject area, some of the earliest studies on summer learning loss focus on the relationship between prior achievement and losses over the summer. Elder (1927) found that students with high reading achievement experience increases in achievement over the summer, while lower readers experience decreases over this time interval. Beggs and Hieronymous (1968) found greater reading achievement losses over the summer months for students with the lowest prior achievement as measured by the Iowa Test of Basic Skills. The same study finds vocabulary losses for lower achieving students and summer gains for higher achieving students. Klibanoff and Haggart (1981), however, did not find a negative relationship between summer achievement loss and initial achievement status. On the contrary, this study found weak evidence that summer loss could actually be more apparent near the top of the achievement scale where regression effects could be at play.

One of the most important implications of summer learning loss is its apparent relationship with poverty. Entwisle and Alexander (1992) popularized the notion that the learning rates of low-income students during the summer months fall behind the learning rates of their wealthier peers. The Copper et al. (1996) meta-analysis confirmed Entwisle and Alexander's (1992) results, finding that while all students, on average, lose mathematics achievement over the summer, reading achievement patterns are economically moderated. Students from lower-income homes tended to lose reading achievement over the summer, while students from higher-income homes stayed the same or gained in reading achievement over the summer months. The authors estimated that this summer learning differential directly results in about a three-month gap between student groups defined by income (Cooper et al, 1996). However, not all studies included in the meta-analysis found such a clear relationship. Ginsburg, Baker, Sweet and Rosenthal (1981) found only a weak relationship between summer achievement change and socioeconomic status. Also, Bryk and Raudenbush (1988) found an opposite, negative relationship where summer losses in mathematics were smaller for the high-poverty schools than the losses observed at the low-poverty schools.

Although empirical support for the finding that poverty is related to summer learning patterns has grown since the Cooper et al. (1996) meta-analysis, a major limitation is that most of the recent studies that have examined this relationship have been done using one, publicly-available dataset. In a review of the literature, McCombs et al. (2011) cite three recent studies that investigate and confirm that high-income students have a summer learning advantage over poorer students, all three of which have used the ECLS-K dataset (Burkam, Ready, Lee, & LoGerfo, 2004; Downey, von Hippel, & Broh, 2004; Benson and Borman, 2010). Additionally, McCoach, O'Connell, Reis, and Levitt (2006) used the same dataset to find that between-school differences in reading achievement can be accounted for, in large part, by differences in summer reading growth over the summer months. While this one dataset is large and of high quality, its limitation is that it can only be used to estimate the summer learning for students who are between kindergarten and first grade. This apparent gap in the literature indicates more research is necessary to explore the persistence of the relationship between summer loss and poverty as students age and progress through the educational system.

## Effect of Summer Learning Loss on MGPs

Summer learning loss is frequently cited as a potential source of bias in growth modeling for teacher and school evaluation (see Haertel, 2013; Larsen, Lipscomb, & Jaquet, 2011;). For example, Papay (2011) found that the impact the summer months has on estimates of teacher effectiveness is substantial, with an observed Spearman rank order correlation between spring-to-spring and fall-to-spring estimates at only $r = 0.7$. Though Papay (2011) suggested summer learning loss is a potential factor contributing to the observed variability in the estimates across the two testing windows, this hypothesis was not formally tested.

A thorough search of the literature revealed four studies that empirically explore effects of summer learning loss on estimates of teacher and school effectiveness. Downey, von Hippel, and Hughes (2008) introduced a new accountability metric called "impact," which explicitly takes into account the variability in summer growth rates. These authors suggested that the difference between the school's average summer growth rate and the average in-school growth rate is a better indicator of school effectiveness than measuring student learning across the year. These authors found non-trivial differences between traditional 12-month growth models and their method which accounts for differences in summer learning. McEachin and Atteberry (2014) used data from the Northwest Evaluation Association's Measures of Academic Progress (MAP) assessments to explore the impact of summer learning loss on measures of school performance. Three research questions structured their inquiry: First, is summer learning loss unevenly distributed across students and schools in a way that leads to systematic bias in aggregated measures of growth? To address this first question, the authors modified growth models by changing the outcome variables to be the achievement gains/losses in the summer months. At least one significant school-effect coefficient in these models indicated bias in the school estimates due to differential summer learning rates which provided evidence for an uneven distribution of summer learning across schools. The study then explores the strength of the relationship between growth estimates and student demographic variables. Results for the second research question showed that across both models and content areas, correlations between percentage of FRL-eligible students in the school and school effectiveness estimates decreased when calculated from fall-to-spring instead of spring-to-spring. Lastly, the authors tested how the ranking of schools based on aggregate measures of growth would change when the testing window moves from spring-to-spring to fall-to-spring. McEachin and Atteberry (2014) found that removing the summer months from the growth estimates increased the likelihood that schools with high percentages of FRL-eligible students are in the upper quintiles of effectiveness.

Both Palardy and Peng (2015) and Gershenson and Hayes (2016) used the ECLS-K data to investigate the effect of summer months on value-added estimated classroom effects. For these studies the authors necessarily limited the full student sample to the randomly chosen subset of students who were tested in the spring of kindergarten, the fall of first grade, and the spring of first grade. Results showed the correlations across the fall-to-spring and spring-to-spring value-added estimates are high, ranging between .77 and .91 depending on the model, subject, and level of aggregation (i.e., teacher- or school-level estimates). The authors argued that even with high correlations such as those observed, the variability in teacher rankings across the estimation periods can result in misclassification for a non-trivial proportion of educators. The authors also found that including student characteristic variables and information related to student summer activities in the spring-to-spring value-added models did not improve the cross-period stability of the measured classroom effects. This is surprising in that the study revealed that the added variability in the teacher estimates due to the summer months could be in part explained by student characteristics, as predicted.

Summer learning loss is a topic that deserves continued study as the landscape of education is shaped by new technology and ever-changing policy. The present study is comprehensive in its scope to build on the limited current literature for SGPs and to contribute to resolving conflicting existing findings. By analyzing data from both content areas and grades 3-8 from two of the largest interim assessment programs in the country, the study builds in a cross-validation of findings.

# DATA

The two studies described in this paper utilize datasets from two widely-adopted interim testing programs: Measures of Academic Progress (MAP) from the Northwest Evaluation Association (NWEA), and STAR from Renaissance Learning, Inc. (RLI). All personally identifiable information had been completely removed from both datasets before coming into the author's possession and both companies have granted authorization for the stated data uses. Both testing programs assess students at multiple time-points throughout the school year using computer adaptive tests that are vertically scaled and aligned to state standards. Both of the datasets have integrated information about FRL eligibility from the National Center for Educational Statistics (NCES) at the school level. Limitations of the datasets include a lack of teacher-student links, meaning we had to create aggregate SGPs at the grade-level within schools in order to investigate the degree of potential bias due to summer loss in the estimates. Thus, the unit of analysis for the current study is often grade level, which refers to all of the students in a given grade within a given school. Depending on the size of the school this could represent a single teacher's class if there is only one teacher per grade, or, it could represent several teachers' classes. Because the correlation between student characteristics and MGPs tends to increase with the size of the sampling unit, aggregating at the grade level rather than at the teacher level may lead to an overestimation of this effect size. Additionally, we do not have information about the academic calendars of the individual schools. Therefore, to estimate summer loss we used the latest measurement occasion from the spring and the earliest measurement occasion for the fall. It is possible that if schools within our datasets have year-round or non-traditional academic years, our methods would be insensitive to these differences in a way that could affect the results of this study. However, because the number of schools following non-traditional academic calendars is likely small, if present in our sample, their influence on the results is likely also to be small.

The MAP dataset contains a large sample from four states for the years 2009-2010. The STAR dataset was drawn from the years 2012-2013, and while smaller, contains data from sixteen states. The states represented in the two datasets are unknown to the authors. Descriptive statistics for the samples are shown in Table 1.

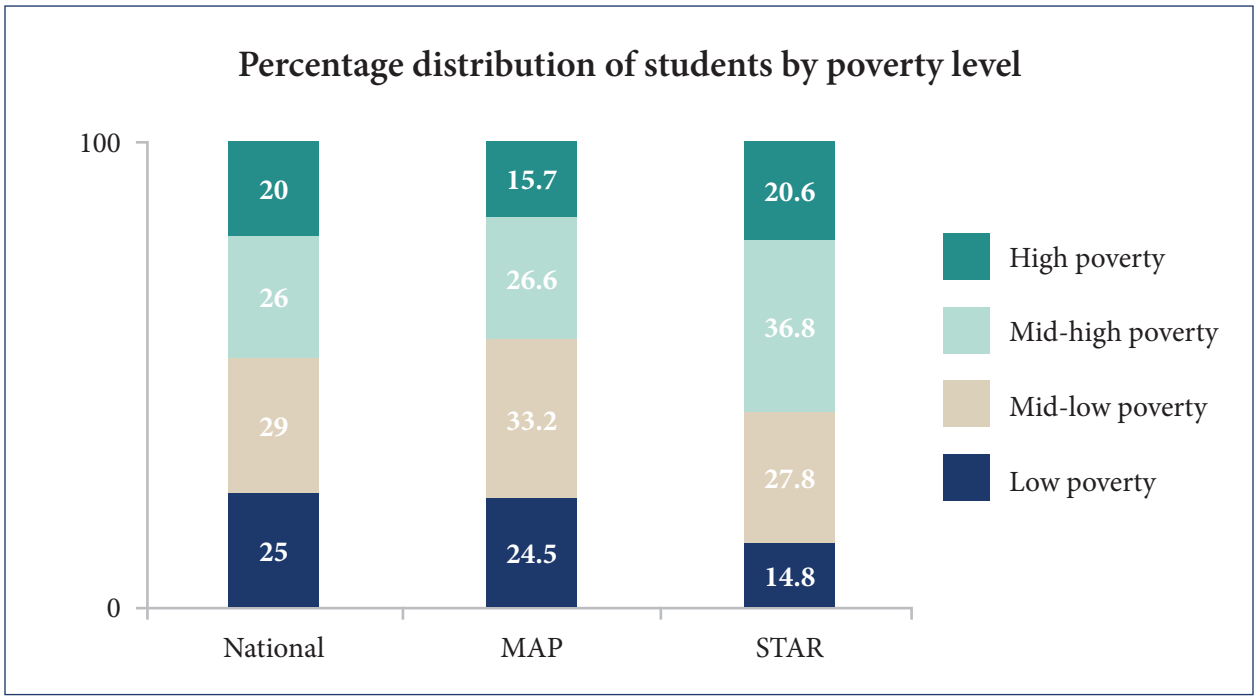
***Table 1: Basic Description of Datasets***

| Testing Program | Prior Spring | Fall | Spring | n States | n students Math | n schools Math | n students ELA | n schools ELA | Grades |
|---|---|---|---|---|---|---|---|---|---|
| MAP | 2009 | 2009 | 2010 | 4 | 14766 | 713 | 14343 | 676 | 4-8 |
| STAR | 2012 | 2012 | 2013 | 16 | 10998 | 89 | 8287 | 90 | 3-8 |

It is clear from the sample sizes presented in Table 1 that the average number of students per school is much greater for the STAR dataset than the MAP dataset. When the grade level minimum sample size is set to 10, the average grade level unit sizes are 21 students with a maximum of 94 for the MAP dataset, and 102 students with a maximum of 420 for the STAR dataset.

Figure 1 shows that the distribution for the percentage of students who are eligible for FRL within the sample schools closely follows the distribution of students nationally (NCES, 2014). The poverty levels are defined by the National Center for Education Statistics where high-poverty schools have more than 75% of the students eligible for FRL, mid-high poverty schools are those with 50.1%-75% FRL, mid-low poverty schools are those where 25.1%-50% of the students are eligible for FRL, and low poverty schools have less than 25% of the students eligible for FRL.

***Figure 1. Percentage Distribution of Schools by Poverty Level. This bar chart compares national distribution of poverty levels to the distributions in the sample datasets.***




For the STAR dataset, SGPs were calculated by the National Center for the Improvement of Educational Assessment (NCIEA) using up to three prior scores. For example, to calculate spring-to-spring SGPs, the 2013 spring score for the was regressed on the 2012 spring score in addition to up to two additional scores for that student occurring prior to the 2012 spring score. For the spring-to-fall SGPs used in this study, the authors calculated the SGPs by regressing the 2012 fall score on the 2012 spring score with no additional priors. For the MAP dataset, SGPs were calculated by the authors using only one prior score for each student. For the spring-to-spring SGPs, the only scores used in the calculation are the 2010 and 2009 spring scores. For the spring-to-fall SGPs, the scores used are the 2009 fall scores and the 2009 spring scores. The nature of the data provided to the authors from NWEA and RLI prevented the option for calculating SGPs using the same number of priors across testing programs. This difference in the methodologies for SGP estimation becomes a major limitation in that the inclusion of multiple prior scores, beyond one previous measurement occasion, has the potential to correct for systematic error introduced by differences in summer learning patterns. This limitation is discussed in more detail in the discussion section of this paper.

Two sets of analyses were conducted to answer the two guiding research questions for this paper. The details of the two studies and their results are provided in the following sections.

# STUDY 1: PREDICTING SUMMER LOSS

In order to estimate the proportion of classroom variance in summer learning patterns that can be accounted for by poverty, SGPs are used to model summer loss. Spring-to-fall SGPs describe individual student movement in the achievement distribution over the summer months. Low spring-to-fall SGPs indicate relative loss, while higher SGPs will indicate relative growth. The purpose of this analysis is to understand any systematic variance in the spring-to-fall grade-level[1] MGPs due to poverty. In order to capture the most power with this analysis, all grades within school units are modeled simultaneously with a series of two-level hierarchical linear models[2]. This multilevel modeling technique is used in order to account for the nested nature of the data where the grade-level sampling units are naturally occurring within schools. The nesting of schools within states is ignored for these analyses.

Model 1: $SF\_SGP_{ij} = \beta_{0j} + \beta_{1j}^{*}(G4_{ij}) + \beta 2_{j}^{*}(G5_{ij}) + \beta 3j^{*}(G6_{ij}) + \beta 4_{j}^{*}(G7_{ij}) + \beta 5_{j}^{*}(G8_{ij}) + e_{ij}$ (3)

$\beta_{0j} = \gamma_{00} + u_{0j}$

Model 2: $SF\_SGP_{ij} = \beta_{0j} + \beta_{1j}^{*}(G4_{ij}) + \beta_{2j}^{*}(G5_{ij}) + \beta_{3j}^{*}(G6_{ij}) + \beta_{4j}^{*}(G7_{ij}) + \beta_{5j}^{*}(G8_{ij}) + e_{ij}$ (4)

$\beta_{0j} = \gamma_{00} + \gamma_{01}^{*}(\%FRL_{j}) + u_{0j}$

where $SF\_SGP_{ij}$ is the spring-to-fall SGPs for student $i$ in school $j$, $\beta_{0j}$ is the MGP at school $j$ in grade 3, and $\beta_{1j}$ to $\beta_{5j}$ are the estimates for the respective dummy-coded grades with grade 3 as the reference group for STAR and grade 4 (since grade 3 information is not available) as the reference group for MAP. The remaining within-school variability in SGPs that cannot be explained by grade is denoted with $e_{ij}$. This level-one equation models the within-school variability in SGPs, while the following level-two equation represents the between-school variability. For Model 1, the baseline model, there are no level-two predictors. In this case, the error term, $u_{0j}$, represents all between-school variability in MGPs. In Model 2, the level-two equation includes the proportion of students eligible for FRL as a predictor for school-level MGPs (%FRL): $\gamma_{00}$ is the average MGP for students in grade three when zero percent of the students in the school are eligible for FRL. The coefficient $\gamma_{01}$ represents the change in $\gamma_{00}$ for a school where 100% of the students are eligible for FRL. A significant, negative coefficient would indicate a significant effect of poverty on summer learning loss. The variance of the error term for this equation, $u_{0j}$, represents the between-school variability in average MGP that cannot be accounted for by %FRL. It is expected that the variance of $u_{0j}$ from Model 2 will be smaller than the variance of $u_{0j}$ from Model 1 if %FRL accounts for any between-school variability in MGPs. A comparison of the two hierarchical linear models estimates the proportion of between-school variance in the spring-to-fall MGPs that can be accounted for by school-level poverty.

Because the degree of relationship between the independent variables and the dependent variable is expected to vary across subject areas, and because not all schools are represented in both mathematics and English Language Arts, the hierarchical linear models are estimated separately for mathematics and ELA for both the MAP and STAR datasets.

[1] Grade levels within schools are used to estimate classroom variance since teacher-student links were not available in the datasets.

[2] The term $\beta_{1j}^{*}(G4)$ is not included in the model for MAP

## Results

The purpose of the first set of analyses is to understand any systematic variance in the spring-to-fall MGPs due to poverty. Table 2 shows the results of the four hierarchical models and reports the coefficient estimates for the %FRL variable, its significance, and the proportion of variance it can account for in between-school variability in MGPs.

***Table 2: %FRL Estimates for Models 1 and 2***

| | n level-1 | n level-2 | $\gamma_{01}$ | *t* ratio+ | *p*-value (one-tailed) | Variance Components | | %level-2 variance accounted for by FRL |
|---|---|---|---|---|---|---|---|---|
| | | | | | | $u_{0j(Baseline)}$ | $u_{0j(FRL)}$ | |
| MAP Math | 14766 | 713 | -7.997 | -4.666 | <0.001 | 52.862 | 50.048 | 5.32% |
| MAP ELA | 14343 | 676 | -5.684 | -3.959 | <0.001 | 20.480 | 18.253 | 10.88% |
| STAR Math | 10998 | 89 | -3.067 | -3.368 | 0.183 | 40.358 | 39.99 | 0.91% |
| STAR ELA | 8287 | 90 | -6.271 | -1.958 | 0.027 | 33.298 | 32.401 | 2.69% |

+ with robust standard errors

Of the four analyses, three of the %FRL coefficients are significant. The significant coefficients range from -5.684 to -7.997, and can be interpreted as the number of MGP points lost in the spring-to-fall MGP scores for schools with 100% FRL eligibility compared to schools with no students eligible for FRL. For MAP math, the proportion of students eligible for FRL accounts for 5.32% of the variability in between-school spring-to-fall MGPs, while for the MAP and STAR ELA samples, this term accounts for 10.88% and 2.69% of the variability, respectively. This finding supports prior research (see Entwisle & Alexander, 1992 and Cooper et al., 1996) that summer loss in reading is economically moderated to a greater extent than in math.

# STUDY 2: IMPROVING THE IMPLICIT MODEL

It is hypothesized that by controlling for summer learning patterns, the observed relationship between MGPs and student characteristics will decrease, resulting in improved estimates of teacher contributions to student growth. This hypothesis stems directly from the forward causal inferences discussed in the introduction, with the hypothesized model—Equation (2). In the context of this analysis, $V_i$ represents spring-to-fall MGPs which are related to both spring-to-spring MGPs, $Y_i$, and the proportion of students in the school who are eligible for FRL, $Z_i$. After controlling for summer learning, $V_i$, the resulting orthogonal—or more likely diminished—relationship between $Y_i$ and $Z_i$ is represented by "$\perp$". Since aggregate MGPs are being examined in these analyses, only units with 10 or more subjects are included in the analysis, which is a common minimum n for many states. The hypothesized model—Equation (2)—is tested by comparing the magnitudes of the following bivariate and partial squared correlation coefficients:

1.) $r^2_{Y,Z}$
2.) $r^2_{Y,Z\,.V}$

Relationship 1 is the squared correlation between spring-to-spring MGPs and the proportion of students eligible for FRL, and relationship 2 is the same squared correlation but after controlling for spring-to-fall MGPs. The difference between the bivariate and partial squared correlations is tested for statistical significance using an *F* test, with an a *priori* type-1 error rate of $\alpha = .05$. The *F* test statistic is calculated in the following way:

$$F = \frac{(r^2_{YZ} - r^2_{YZ.V})/1}{(1 - r^2_{YZ})/(n-2)} \qquad (5)$$

This ratio represents the percentage of residual variance from the bivariate correlation that can be accounted for by controlling for poverty. This *F* statistic formula was derived based on the statistic used to test the difference between two, multiple $R^2$ values (as seen in Cohen, Cohen, West, and Aiken, 2003, p. 171), but rather than comparing variance accounted for by multiple $R^2$ values, the variance components being compared are associated with the bivariate and partial correlation coefficients.

## Results

Tables 3 and 4 show the results for MAP and STAR, respectively, where the grade level within schools is the unit of analysis. Grades 7 and 8 in mathematics and grades 6-8 in ELA were not analyzed in the STAR dataset due to small sample sizes.

***Table 3: Results for MAP Datasets with Grade as unit of analysis (n ≥ 10)***

| MATH | n | $r_{Y,Z}$ | $r^2_{Y,Z}$ | $r^2_{Y,Z\,.V}$ | *F* | p-value |
|---|---|---|---|---|---|---|
| Grade 4 | 376 | -0.194 | 0.038 | 0.009 | 11.358 | <.001 |
| Grade 5 | 359 | -0.157 | 0.025 | 0.006 | 6.860 | 0.005 |
| Grade 6 | 234 | -0.035 | 0.001 | 0.000 | 0.271 | 0.668 |
| Grade 7 | 207 | -0.097 | 0.009 | 0.004 | 1.209 | 0.198 |
| Grade 8 | 188 | -0.115 | 0.013 | 0.008 | 1.057 | 0.228 |

| ELA | n | $r_{Y,Z}$ | $r^2_{Y,Z}$ | $r^2_{Y,Z.V}$ | *F* | p-value |
|---|---|---|---|---|---|---|
| Grade 4 | 364 | -0.174 | 0.030 | 0.027 | 1.307 | 0.181 |
| Grade 5 | 355 | -0.080 | 0.006 | 0.005 | 0.334 | 0.583 |
| Grade 6 | 235 | -0.122 | 0.015 | 0.005 | 2.287 | 0.084 |
| Grade 7 | 199 | -0.066 | 0.004 | 0.002 | 0.458 | 0.468 |
| Grade 8 | 183 | -0.050 | 0.003 | 0.002 | 0.087 | >.999 |

***Table 4: Results for STAR Datasets with Grade as unit of analysis (n ≥ 10)***

| MATH | n | $r_{Y,Z}$ | $r^2_{Y,Z}$ | $r^2_{Y,Z.V}$ | *F* | p-value |
|---|---|---|---|---|---|---|
| Grade 3 | 46 | -0.313 | 0.098 | 0.091 | 0.351 | 0.560 |
| Grade 4 | 52 | -0.181 | 0.033 | 0.035 | | |
| Grade 5 | 40 | 0.087 | 0.008 | 0.024 | | |
| Grade 6 | 34 | 0.125 | 0.016 | 0.011 | 0.155 | 0.929 |
| Grade 3 | 46 | -0.313 | 0.098 | 0.091 | 0.351 | 0.560 |
| **ELA** | **n** | $r_{Y,Z}$ | $r^2_{Y,Z}$ | $r^2_{Y,Z.V}$ | *F* | p-value |
| Grade 3 | 34 | -0.102 | 0.010 | 0.002 | 0.284 | 0.642 |
| Grade 4 | 43 | -0.064 | 0.004 | 0.002 | 0.083 | >.999 |
| Grade 5 | 39 | -0.009 | 0.000 | 0.001 | | |

As shown in Tables 3 and 4, significant differences between the bivariate and partial squared correlations occur at grades 4 and 5 in mathematics for the MAP datasets. This means that after controlling for spring-to-fall MGPs, the correlation between grade-level spring-to-spring MGPs and the proportion of students within the grade eligible for FRL decreased significantly. Student poverty went from accounting for 3.8% and 2.5% of the variance in spring-to-spring MGPs for grades 4 and 5 respectively, to only .9% and .6%. This is an indicator of statistically significant bias in the spring-to-spring estimates for the lower grades in the mathematics MAP dataset. The changes in the correlations are shown in Figures 2 and 3.

***Figure 2. Changes in correlation magnitude for Grade 4 Math – MAP Dataset***

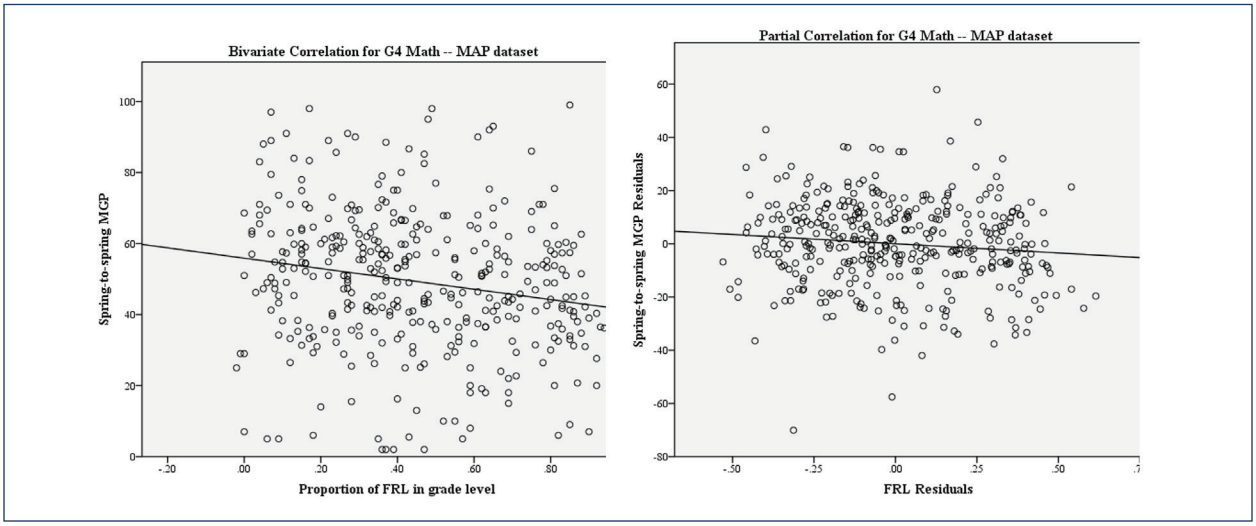
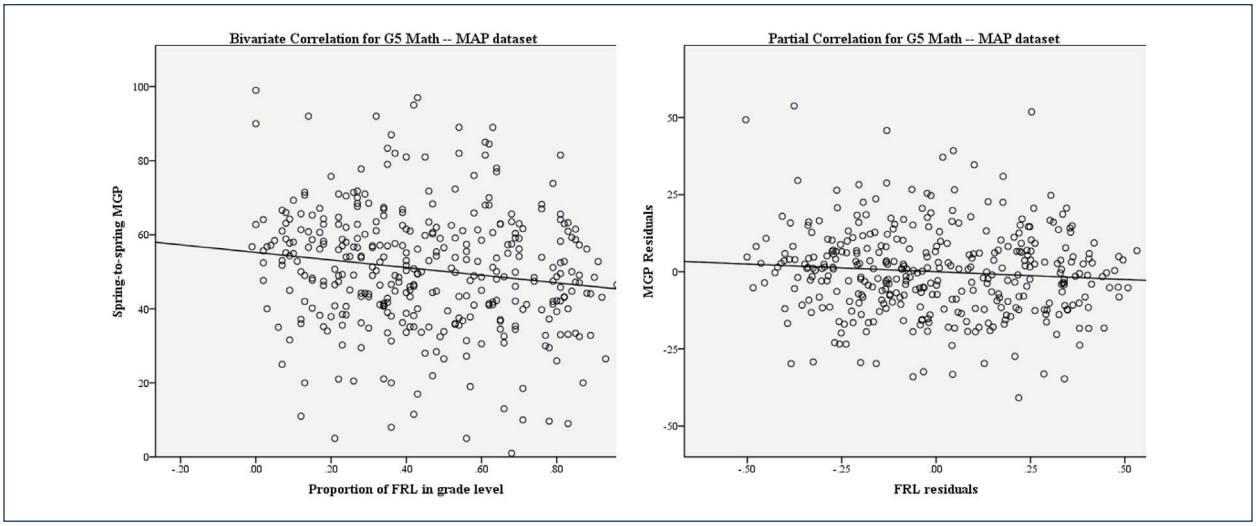


*Figure 3. Changes in correlation magnitude for Grade 5 Math – MAP Dataset*




The rest of the correlational analyses revealed that controlling for the summer months did not significantly reduce the correlation between spring-to-spring MGPs and poverty.[3] In three cases, STAR grades 4 and 5 mathematics and grade 5 in ELA, controlling for the summer months actually increased the relationship between spring-to-spring MGPs and poverty. These results indicate that with exception of grades 4 and 5 in the MAP mathematics dataset, the observed correlation between spring-to-spring grade-level MGPs and poverty is not due to systematic differences in learning patterns over the summer months. In general, the STAR datasets do not produce any significant findings. This does not seem to be due to a lack of power to detect an effect; rather, the effect sizes themselves are smaller for the STAR dataset. The SGPs in the STAR dataset are less correlated with poverty to begin with. This is a positive finding and is likely due to the fact that the STAR SGPs take into account multiple priors (which was not able to be done with the MAP SGPs given the nature of our dataset). This suggests that including more than one prior score in the calculation of SGPs may provide a protective effect against any bias introduced by differential summer learning patterns by accounting for student learning differences across multiple time periods.

[3] When changes in the magnitude of correlations happen in the unexpected direction, F-values are negative and are therefore not provided in the tables of results.

# DISCUSSION

The primary issue this research attempts to better understand is the known relationship between aggregate measures of student growth and student characteristics such as poverty. The guiding hypothesis was that economically-moderated summer learning patterns are in part driving this correlation, which if so, represents bias in the growth estimates when used for teacher or school evaluation. Using data from two, national interim testing programs, this hypothesis was investigated by addressing two research questions:

1. What proportion of variance in summer learning patterns can be accounted for by poverty? And,
2. Does controlling for loss over the summer months reduce the magnitude of the relationship between MGPs and student-level poverty?

Spring-to-fall SGPs were calculated and analyzed to understand normative summer loss using a series of hierarchical linear models. The results showed that school-level poverty was a significant predictor of aggregated grade-level MGPs for both mathematics and ELA in the MAP dataset and for ELA in the STAR dataset. This means that, in some cases, differences in summer learning patterns can be systematically explained by the level of poverty of the students within the school. There was not a statistically significant relationship ($p > .05$) between poverty and spring-to-fall MGPs for mathematics in the STAR dataset. In the significant analyses, the percentage of between-school variability in MGPs accounted for by the FRL variable ranged from 2.69% in STAR ELA to 10.88% in MAP ELA. This means that though poverty can explain some between-school variation, the majority of the differences in summer learning patterns remained unexplained. While informative, this analysis did not detect how much influence the explained portions of variance may have on annual estimates of MGPs. Research question 2 investigates the influence of the summer months on the correlation between spring-to-spring MGPs and the proportion of students eligible for FRL.

To test the influence of the summer months on the observed correlation between MGPs and FRL, a series of bivariate and partial correlations were evaluated and their differences tested. Of all the analyses, only grades 4 and 5 for MAP mathematics showed a significant reduction in the correlations between spring-to-spring MGPs and FRL. This means that when the grade-level is the unit of analysis, bias in MGPs due to summer learning loss may be a concerning issue in mathematics grades. Though the correlations between spring-to-spring MGPs and %FRL are small in magnitude to start with, any reduction in the correlation represents a degree of bias due to the summer months, which are typically out of the school's control. The significant reduction in shared variance between MGPs and %FRL indicates that using the spring-to-spring MGPs for educator evaluation, when calculated the way they had been for the MAP dataset using only a single prior, may result in a misspecification of within-year growth, with a downward bias for those grades or classrooms with higher proportions of students eligible for FRL. This means that variability in student learning over the summer months may be contributing to the phenomenon that educators serving more disadvantaged students have, on average, lower MGPs. Because of this bias in the single-prior MGPs, the metric may be not be a fair estimate of educator effectiveness.

Interestingly, the STAR dataset showed no significant reduction in the correlations. The MGPs calculated for STAR seem to be less influenced by poverty, as the amount of shared variance between spring-to-spring MGPs and FRL to start with was much smaller than for MAP across the grade levels. This difference may

be an artifact of the way the SGPs were calculated, using multiple prior achievement indices, rather than conditioning on a single prior score, as was done for the MAP dataset. This is good news in that bias due to systematic differences in summer learning patterns appears to be mitigated by including multiple prior years of data.

Due to a lack of reliable links between students and teachers, this study analyzes variance in MGPs at the grade level. While this limits generalizability, it is likely that any systematic variance in grade MGPs would also occur at the classroom level, to a somewhat lesser extent. Therefore, the findings of this study serve as a framework for understanding the effects of summer learning loss on MGPs and warrant further research at the classroom level.

# CONCLUSIONS AND POLICY IMPLICATIONS

Does systematic variance in summer learning loss contribute to bias in annual estimates of student growth for school personnel evaluation? The answer is conditionally yes, and the degree to which this is a real issue
varies across testing programs, grades, and subject areas. Moderate correlations between spring-to-spring SGPs and summer loss suggest that variations in summer learning patterns may influence annual estimates of student growth. However, because the summer loss that was detected in our datasets does not seem to be primarily a function of student poverty, simply controlling for student poverty will not likely significantly alleviate the issue. Poverty only explains between zero to 11% of the between-school variability in summer learning patterns as measured by spring-to-fall MGPs. Based on the strength of the correlations between annual MGPs and summer loss, our first policy recommendation is that when designing student growth models to be used for teacher evaluation, it does not make as much sense to control for variables that may affect summer growth patterns (e.g., as is done with student poverty in some value-added models) than more directly controlling for the differential summer patterns themselves. The results of this study show that while poverty is a significant factor, it is only one, relatively small factor that can explain the influence of the summer months on MGPs. Instead of controlling for student poverty in the model, a more effective way of reducing any bias that may be introduced by summer learning loss would be to control directly for the summer months. Given new flexibility in assessment systems offered under the Every Student Succeeds Act, states may want to consider how the use of multiple assessments throughout the school year could be used to track growth more effectively than a single annual assessment. Future research is needed to further investigate whether growth estimates calculated for the academic school year (i.e., fall-to-spring—with multiple priors), rather than based on annual measurements, would lead to more valid estimates of student growth for educator and school evaluation.

Secondly and unsurprisingly, the number of prior observations makes a significant difference when calculating SGPs. For the STAR dataset, controlling for variability in the summer months did not significantly decrease the relationship between spring-to-spring MGPs and poverty. This may be because the STAR SGPs were calculated using more than one prior achievement score, which can account for the variability in summer learning patterns across time. If a student who loses achievement over the summer months one year is more likely to lose the next year, then using multiple years of data may lead to more accurate determinations of normative student growth. Though this conclusion has previously been shown using simulated data (see Castellano & Ho, 2013), we do believe it is worth re-emphasis as it stresses the real-world importance of maintaining and then utilizing large, longitudinal datasets to strengthen the validity of SGPs and growth models in general.

In sum, study results show that systematic differences in summer learning do not seem to be the driving factor in the correlation between poverty and MGPs. This is in part explained by the finding that summer learning patterns are not primarily a function of poverty. Therefore, the implicit model—Equation (1)—should be reconsidered as a likely explanation for the observed correlations between MGPs and student poverty. The findings of this paper suggest that the correlations between poverty and MGPs cannot be fully explained by omitted variable bias, but may be instead a result of an inequitable distribution of teacher quality. This conclusion warrants further study as the policy implications for addressing a systematically uneven distribution of effective teachers across school settings has the potential for better understanding and reducing the persistent achievement gaps in the United States.

# REFERENCES


Beggs, D. L., & Hieronymus, A. N. (1968). Uniformity of growth in the basic skills throughout the school year and during the summer. *Journal of Educational Measurement, 5*(2), 91-97.

Benson, J., & Borman, G. (2010). Family, neighborhood, and school settings across seasons:

When do socioeconomic context and racial composition matter for the reading achievement growth of young children. *Teachers College Record, 112*(5), 1338-1390.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155–170). New York: Taylor & Francis.

Betebenner, D., VanIwaarden, A., Domingue, B., & Shang, Y. (2016). SGP: Student Growth Percentiles & Percentile Growth Trajectories. (R package version 1.5-0.0) URL: sgp.io

Betts, J. R., Zau, A., & Rice, L. (2003). *Determinants of student achievement: New evidence from San Diego* (pp. 1-5821). San Francisco: Public Policy Institute of California.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). The draw of home: How teachers' preferences for proximity disadvantage urban schools. *Journal of Policy Analysis and Management, 24*(1), 113-132.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop.* National Academies Press.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108.

Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education, 77*(1), 1-31.

Castellano, K. E., & Ho, A. D. (2012). Simple Choices among Aggregate-Level Conditional Status Metrics: From Median Student Growth Percentiles to Value-Added Models. *Unpublished manuscript.*

Castellano, K.E., & Ho, A. D. (2013). Contrasting OLS and Quantile Regression Approaches to Student "Growth" Percentiles. *Journal of Educational and Behavioral Statistics, 38*(2), 190-215.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers? *Journal of Policy Analysis and Management, 23*(2), 251-271.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. (1966). *Equality of educational opportunity.* Washington, DC: Department of Health, Education and Welfare.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227–268.

Darling-Hammond, L. (1995). Inequality and access to knowledge. In J. A. Banks (Ed.), *The Handbook of Research on Multicultural Education.* New York: Macmillan.

Darling-Hammond, L. (1996). What matters most: A competent teacher for every child. *Phi Delta Kappan, 78*(3), 193-200.

Downey, D. B., Von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review, 69*(5), 613-635.

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.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2013). *Selecting Growth Measures for School and Teacher Evaluations: Should Proportionality Matter?* CALDER Working Paper No. 80.

Ehlert, M., Koedel, C., & Podgursky, M. (2014, Spring). Choosing the right growth measure. *Education Next, 14*(2). Retrieved from http://educationnext.org/choosing-the-right-growth-measure/

Entwisle, D. R., & Alexander, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review*, 72-84.

Gelman, A. (2011). Causality and statistical learning. *American Journal of Sociology, 117*(3), 955-966.

Gelman, A., & Imbens, G. (2013). *Why ask why? Forward causal inference and reverse causal questions* (No. w19614). National Bureau of Economic Research.

Gershenson, S., & Hayes, M. (2016). The implications of summer learning loss for value-added estimates of teacher effectiveness. *Educational Policy, 30*(4).

Ginsburg, A., Baker, K., Sweet, D., & Rosenthal, A. (1981, April). *Summer learning and the effects of schooling: A replication of Heyns.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28-39.

Guarino, C. M., Santibañez, L., & Daley, G. A. (2006). Teacher recruitment and retention: A review of the recent empirical literature. *Review of Educational Research, 76*(2), 173-208.

Haertel, E. (2013, March 21). Reliability and validity of inferences about teachers based on student test scores. *14th William H. Angoff Memorial Lecture.* Lecture conducted from ETS in Princeton, New Jersey. Retrieved from http://atlanticresearchpartners.org/wp-content/uploads/2013/07/PICANG14.pdf

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326-354.

Harris, D., Sass, T. (2009). What makes for a good teacher and who can tell? CALDER Working Paper No. 30. *Urban Institute.*

Helf, S., Konrad, M., & Algozzine, B. (2008). Recouping and rethinking the effects of summer vacation on reading achievement. *Journal of Research in Reading, 31*(4), 420-428.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700-709.

Jacob, B. A. (2007). The challenges of staffing urban schools with effective teachers. *The Future of Children, 17*(1), 129-153.

Klibanoff, L. S., & Haggart, S. A. (1981). *Summer growth and the effectiveness of summer school.* System Development Corporation.

Ladson-Billings, G. & Tate, W. F. (1995). Toward a critical race theory of education. *Teachers College Record*, 97, 47–68.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

Larsen, S. E., Lipscomb, S., & Jaquet, K. (2011). *Improving school accountability in California.* Public Policy Institute of California.

Lissitz, R. W. (2012, April). *The evaluation of teacher and school effectiveness using growth models and value added modeling: Hope versus reality.* Presented at the Annual Meeting of the American Educational Research Association, Vancouver, BC.

Marland, J. (2014, July). *Conceptualizing and Analyzing the Relationship Between Prior Achievement and Growth.* Presented at the National Center for the Improvement of Educational Assessment.

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice, 34*(1), 15-21.

McCall, M. S., Hauser, C., Cronin, J., Kingsbury, G. G., & Houser, R. (2006). Achievement Gaps: An Examination of Differences in Student Achievement and Growth. The Full Report. *Northwest Evaluation Association.*

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98*(1), 14.

McCombs, J. S., Augustine, C. H., Schwartz, H. L., Bodilly, S. J., McInnis, B., Lichter, D. S., & Cross, A. B. (2011). *Making summer count. How summer programs can boost children's learning.* Santa Monica, CA: The RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/monographs/2011/RAND_MG1120.pdf

McEachin, A. & Atteberry, A. (2014) The Impact of Summer Learning Loss on Measures of School Performance. *EdPolicyWorks Working Paper Series, No. 26.* Retrieved from: http://curry.virginia.edu/uploads/resourceLibrary/26_McEachin_Summer_Learning_Loss.pdf

National Center for Education Statistics. (2014, April). *Concentration of public school students eligible for free or reduced-price lunch.* Retrieved from http://nces.ed.gov/programs/coe/indicator_clb.asp

NCES, see National Center for Education Statistics.

Oakes, J., & Lipton, M. (1993). Tracking and ability grouping: A structural barrier to access and achievement. In Bellanca, J., & Swartz, E. (Eds.). *The Challenge of Detracking: A Collection.* Skylight Publishing: Palatine, IL.

Palardy, G. J., & Peng, L. (2015). The effects of including summer on value-added assessments of teachers and schools. *Education Policy Analysis Archives, 23,* 92.

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.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raudenbush, S., Bryk, A., & Congdon, R. (2013). HLM 7.01 for Windows [Hierarchical linear and nonlinear modeling software].

Sandberg Patton, K. L., & Reschly, A. L. (2013). Using curriculum-based measurement to examine summer learning loss. *Psychology in the Schools, 50*(7), 738-753.

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for Student Growth Percentiles using the SIMEX method. *Educational Measurement: Issues and Practice, 34*(1), 4-14.

Sireci, S. G., Wells, C. S., Bahry, L. (2013, April). *Student growth percentiles: More noise than signal?* Unpublished paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

U.S. Department of Education. (2009, January 8). *Secretary Spellings Approves Additional Growth Model Pilots for 2008-2009 School Year.* Retrieved from http://www2.ed.gov/news/pressreleases/2009/01/01082009a.html

U.S. Department of Education. (2012, June 7). *ESEA Flexibility.* Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

U.S. Department of Education. (2013, June 20). *Fact Sheet: Accountability Timeline for New College- and Career-Ready Standards.* Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/assessment-transition/fact-sheet.doc

Wright, P. S. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS Institute Inc: http://www.sas.com/resources/whitepaper/wp_16975.pdf


susan@lyonsassessment.com
www.lyonsassessmentconsulting.com