Effect of Summer Learning Loss on Aggregate Estimates of Student Growth

By

Susan Connolly Gillmor

Submitted to the graduate degree program in Psychology and Research in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson John Poggio, Ph.D.

William Skorupski, Ed.D.

Vicki Peyton, Ph.D.

Marianne Perie. Ph.D.

Marc Mahlios, Ph.D.

David Hansen, Ph.D.

Date Defended: April 29, 2015

The Dissertation Committee for Susan Connolly Gillmor
certifies that this is the approved version of the following dissertation:

Effect of Summer Learning Loss on Aggregate Estimates of Student Growth

Chairperson John Poggio, Ph.D.

Date approved: April 29, 2015

An Abstract of

Effect of Summer Learning Loss on Aggregate Estimates of Student Growth

By

Susan Connolly Gillmor

Submitted to the Graduate Faculty in partial fulfillment of the requirements for the Doctor of Philosophy Degree.

University of Kansas
April 2015

Recent reforms in federal educational policy now mandate the use of student assessment data to evaluate teachers and principals. Despite the widespread adoption of Student Growth Percentiles (SGPs) and other models to link student achievement growth to teacher and school effectiveness, little research exists evaluating the validity of the resulting effectiveness estimates for use in high-stakes personnel evaluation systems. This paucity in the literature is especially problematic given that significant correlations between effectiveness estimates and student characteristics, specifically poverty, have been well documented. This dissertation explores summer learning loss as one potential source of bias in annual estimates of student growth for teacher and school evaluation. The guiding hypothesis is that economically moderated summer learning patterns are contributing to systematic error variance in teacher and school effectiveness estimates when calculated based on annual test scores. Datasets from two, nationally distributed commercial interim assessment programs are analyzed separately and their results discussed. Results reveal that the extent of summer learning loss, and by extension, its effect on the validity SGPs for evaluation purposes varies by subject, grade level, and testing program. Statistically significant correlations between mean Student Growth Percentiles and summer learning loss range from $r = -.310$ to $r = -.662$. Implications for fairness and education policy are discussed.

For Sally McGoey, Susan Plunket, Tessa Villaseñor Zapata and all public schools teachers who continue to raise the bar and bring their students with them.

Acknowledgements

This dissertation and my success in this degree program would not have been possible without the love, support, and guidance I received from my immediate and extended families and friends. Thank you to Bill Lyons, Linda Plunket, Reg Gillmor, Michael Gillmor, and Bridgette Donahue for your much needed words of encouragement throughout this entire journey. I am so blessed to have a network of so many smart and kind people who are always available to listen, discuss, and push me along. This degree is the result of a group effort, and I am so grateful for your faith in me.

Thank you to all of my professors in the REMS program at the University of Kansas, Carol Woods, and to my dedicated committee members. John Poggio, Billy Skorupski, Vicki Peyton, Marianne Perie, and Marc Mahlios, you have encouraged me to explore my interests and given me the support I needed to write a dissertation that I am proud of. I am grateful to have found such a generous group of scholars, thank you for your time and your energy.

Thank you to Scott Seider for providing the best introduction to educational research an early scholar could have. I am grateful for your encouragement and including me in such high quality and interesting work. The undergraduate summer research experience, under your direction, has greatly influenced the trajectory of my academic and professional career. Thank you for being a role model and mentor.

Thank you to Scott Marion, Damian Betebenner, Charlie DePascale and all the staff at the Center for Assessment for their guidance, conversation, and support in conceptualizing and finishing this paper. I have very much enjoyed working with you this year and I am grateful to be beginning my career in an environment with such expertise and intellectual curiosity.

A very special thank you to my advisor, mentor, and friend John Poggio. Thank you for the many, many hours you have dedicated to my education, teaching me not only about statistics, evaluation, and measurement, but about what it means to be a professional, a scholar, and a teacher. I remember the very first time I went to your office hours, I was confused about a question you posed in your Regression Analysis homework about the theoretical upper limit for the standard error of the mean. The following discussion started a conversation that, from that day forward, has never ended. Thank you for keeping the conversation going, and in doing so, transferring your passion for the content to me. I am so grateful to have found such a deep source of knowledge and encouragement in you. Thank you for your commitment to me and my education.

Lastly, thank you to Nate Jensen at the Northwest Evaluation Association, and Laurie Borkon at Renaissance Learning for your commitment to supporting independent research. It has been a pleasure working with both of you and I appreciate the time and effort you have dedicated to serving academic research.

Table of Contents

Introduction

Federal policy has recently mandated that states use student growth data to evaluate teachers (U.S. Department of Education, 2012). Because states did not reach the achievement target of 100% proficiency by 2014, as prescribed by the 2001 re-authorization of the Elementary and Secondary Education Act (ESEA), the current administration implemented a flexibility waiver program. In exchange for relaxing the 100% proficiency mandate, flexibility waivers require states to implement a number of new accountability reforms including a strong emphasis on using student growth data for school personnel evaluation purposes. The ESEA flexibility waivers require 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). The waiver renewal process guidelines released August 30, 2013 continue this emphasis and further require the implementation of these high-stakes personnel decisions by the 2014-2015 school year (U.S. Department of Education, 2013). For a variety of political and technical reasons, recent communications from the U.S. Department of Education have eased away from the initial fast-paced timeline for implementing the evaluation reforms and now allow states to apply for a delay in the full implementation of student growth-based evaluation systems to the 2015-2016 school year (Camera, 2014). Even with this extension, states are continuing to revise their teacher evaluation systems rapidly to include links to student achievement growth as measured by standardized tests.

One method of measuring student achievement growth for the purpose of educator evaluation is the Student Growth Percentile (SGP) model (Betebenner, 2008). The SGP model uses quantile regression (Koenker, 2005) to rank student growth relative to peers with the same prior achievement scores. SGPs are aggregated at the teacher and school levels using the mean or median of all student scores to result in a single summary score for the aggregated unit: the mean/median Student Growth Percentile (MGP). Recent research has shown that mean-based SGPs have more desirable statistical properties over median SGPs (Castellano & Ho, 2012). Unlike some of the other growth models used for teacher evaluation, such as many Value-Added Models (VAMs), SGPs condition only on prior achievement and leave out other student-level variables such as demographic and socioeconomic information. Additionally, SGPs are generally considered to be a description of how much students have grown relative to peers with similar prior achievement, and are not designed to measure teacher effectiveness or isolate a portion of student growth that can be directly attributable to the teacher (Betebenner, 2008). In this sense, unlike many VAMs, teacher effectiveness estimates derived from the Student Growth Percentile model, MGPs are designed to be a descriptive indicator for use within a comprehensive evaluation system rather than a direct measure of teacher quality. This means that, by design, MGPs should be considered within the context of the teaching environment and in relation to other indicators of teacher quality rather than as a quantitative measure to be used as part of a "total score" for teacher effectiveness. Though MGPs are operationalized quite differently than VAMs, the resultant estimates tend to correlate highly (Briggs & Betebenner, 2009; Goldhaber, Walch, & Gabele, 2014; Ehlert, Koedel, Parsons, & Podgursky, 2013).

Student Growth Percentiles have become the most prevalent 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 popularity, 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 effectiveness estimates to correlate with student characteristics. Goldhaber, Walch, and Gabele (2014) warn that differences in the make-up of students can have substantial effects on teacher-level MGPs. These authors find that as the average prior achievement level of the classroom increases by one standard deviation, MGPs increase by an average of 15 and 25 percentile points for math and reading, respectively. Wright (2010) finds statistically significant correlations from -0.13 to -0.31 between teacher-level MGPs and percentage of students eligible for free/reduced lunch in the classroom. At the school level, Ehlert, Koedel, Parsons, and Podgursky (2013) find that disadvantaged schools are disproportionally underrepresented in the top quartile of schools on the median Student Growth Percentile metric. The underlying causes of the observed correlations, however, are unknown. The extent to which these correlations reflect the reality of uneven distribution of teacher quality, or are rather, manifestations of model bias is not yet well understood.

Defining the Problem

Gelman and Imbens (2013) argue that research on 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 potential the 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 a way to formalize scientific inquiry as a model-building process. In the context of the current research, the implicit model would be that teacher effectiveness estimates, as derived from mean Student Growth Percentiles, are correlated with student-level characteristics. Gelman and Imbens (2013) encourage researchers to question this relationship and ask "why?" What could be causing the observed correlation? The purpose of posing this question is not to estimate parameters or confidence intervals, but rather to improve the implicit model—to identify potential sources of confounding variables.

Because Student Growth Percentiles rank students only relative to their peers with similar prior achievement, the effect of the modeling results in a zero relationship between measured student growth and the student's previous achievement. In other words, no matter where students lie on the achievement continuum, they have equal or near equal probabilities of obtaining all possible growth scores. Therefore, it cannot be that the observed relationship at the aggregate, teacher level is due to the presumably different absolute growth trajectories at different points along the achievement scale (e.g., "Matthew Effects." see Walberg & Tsai, 1983). Given this, two alternative models are possible as adapted from Gelman and Imbens (2013):

$$Y_i(x) \perp Z_i, \text{ and} \tag{1}$$

$$Y_i \perp Z_i \mid V_i \tag{2}$$

where, in the context of the current study, Y_i is an individual teacher MGP, and Z_i represents the student-level characteristics in a classroom. Let $Y_i(x)$ represent all possible outcomes in the outcome space, each of which denoted as $Y_i(X)$. In model 1, the observed association between the teacher 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 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 might be a function of district wealth, X_i. 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 the causal inference framework introduced by Gelman and Imbens (2013), two, not mutually exclusive, possible causes for the observed relationship between average prior achievement and MGPs are discussed: 1) an uneven distribution of teacher quality, and 2) a manifestation of omitted-variable bias in the model. The first possible cause of the observed relationship between teacher- and school-level MGPs and average prior achievement 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 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 means 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 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 low-achieving schools contributes to an overall decline in the average quality of the teacher workforce at these schools and therefore likely leads to an uneven distribution of teacher quality across all schools (Clotfelter, Ladd, Vigdor, & Diaz, 2004; Darling-Hammond, 1995).

While it is now culturally accepted that student achievement is truly correlated with background factors of students, and the observed relationship is not a manifestation of test bias alone (Coleman et al., 1966), the same cannot be said for the correlation between teacher effectiveness estimates and student background variables. An alternative possibility for explaining this relationship is that omitted, confounding variables related to both teacher effectiveness estimates and student-level characteristics could be 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 who have the same true effectiveness—who are equally effective at eliciting achievement growth from their students—receive different MGP estimates due to factors outside of their control (e.g., classroom demographic characteristics). 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 likely 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 may not only influence a student's achievement status, but also impact 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 a design standpoint, it is vital to disambiguate 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, overcompensate 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 the hopes of improving 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 achievement in the summer months (Entwisle & Alexander, 1992). Although this pattern has been documented extensively by educational 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. These annual estimations include the summer months over which educators have little-to-no control. The teacher effectiveness estimates are essentially absorbing the positive or negative influences the summer vacation period has on their students' achievement. Written even before the latest accountability movement, 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 guiding hypothesis for this study is that conditioning on prior achievement, as is done in the Student Growth Percentile model, may not be sufficient for capturing the different, and likely economically-moderated, growth rates that occur during the summer months. Because Student Growth Percentile estimates are typically calculated annually, summer learning loss is hypothesized to be correlated with not only student characteristics but also with MGPs. Accounting for the summer months, might therefore improve the implicit model with the following hypothesized model:

$$Y_i \perp Z_i \mid V_i \tag{2}$$

where V_i represents summer learning loss that is related to both MGPs, Y_i, and student characteristics, Z_i, and $\perp$ represents an orthogonal or diminished relationship. To test this model, three research questions structure the study:

1. What is the effect size of summer learning loss? How do average summer losses compare by subject, grade-level, prior achievement, and poverty-level?
2. What proportion of variance in summer learning patterns can be accounted for by poverty?

3. Does controlling for loss over the summer months reduce the magnitude of the relationship between mean Student Growth Percentiles (MGPs) and student-level poverty?

Literature Review

Student Growth Percentiles

The Student Growth Percentile model was originally created for the Colorado State Department of Education by Betebenner 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 relative to a peer group. 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:

$$\text{Student Growth Percentile} = P\,(\text{Current Achievement} \mid \text{Past Achievement}) * 100 \qquad (3)$$

This probability is calculated by estimating the probability density of the student's current score after conditioning on the student's prior score(s). In the Student Growth Percentile model, conditional densities are estimated with quantile regression (Koenker, 2005). Quantile regression is comparable to ordinary least squares regression but is median-based rather than mean-based. The conditional quantile functions used to estimate Student Growth Percentiles are specified using R (R Development Core Team, 2014) with the SGP package (Betebenner, 2009). Mean Student Growth Percentiles (MGPs) are calculated by averaging the Student Growth Percentiles of all of the students in the unit (e.g., classroom, grade level, 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.

This study uses the Student Growth Percentile 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 race and poverty in addition to prior achievement when estimating teacher effects. Because the Student Growth Percentile 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 (Ehlert, Koedel, Parsons, & Podgursky, 2014). Secondly, the SGP model continues to grow in popularity and has been adopted by more than half of the states for use in teacher evaluation systems. 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 59 results, while the same database returns only three 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 achievement growth rates are moderated by subject matter, grade level, prior achievement, and poverty. Tests in math and literacy reveal different summer growth patterns. A 1996 meta-analysis of 13 studies finds that summer learning loss tends to be more dramatic in mathematics than in reading (Cooper et al., 1996) with respective effect sizes of d = -.14 and d = -.05. This could be because families more often spend more time reading at home than promoting or practicing math skills (Harris & Sass, 2009). More recently, Helf, Konrad, and Algozzine (2008) find no evidence of summer setback for reading as they do in math, and instead report summer gains, especially for those students towards the bottom of the achievement scale. On the contrary, Burkam, Ready, Lee and LoGerfo (2004) find evidence of math gains over the summer months, with gains being highest for higher-income students. This 2004 study relies on data from the Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K), to calculate summer learning from kindergarten to first grade, which means these findings may not be generalizable across all grade levels.

Some studies suggest that there is a relationship between grade level and the amount of summer loss, but results about the directionality of the relationship are mixed. The Copper et al. (1996) meta-analysis finds a positive relationship between grade level and summer learning loss, which means that as students progress through school, summer learning loss becomes more apparent. However, more recent studies report the opposite; both Borman and Dowling (2006) and Sandberg Patton and Reschly (2013) find the summer learning loss to be greater at lower grade levels. The Borman and Dowling (2006) study finds greater loss in kindergarten than first grade, while similarly, the Sandberg Patton and Reschly (2013) study finds greater losses at grades two and three than in grades four and five. As a qualification to the results for the Sandberg Patton and Reschly (2013) study, while loss for grades two and three was statistically significant, the effect size was small—approximately 20%-30% of a standard deviation. Though these are small, they are similar in size to other research in the field (Allinder & Eicher, 1994). Neither of these studies examined higher grade levels where the summer loss may again increase as suggested by the Cooper et al. (1996) meta-analysis.

In addition to being moderated by subject area and grade level, some of the earliest studies on summer learning loss focus on the relationship between achievement and summer loss. Elder (1927) finds 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) find 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, do not find a negative relationship between summer achievement loss and initial achievement status. On the contrary, this study finds weak evidence that summer loss could actually be more apparent near the top of the achievement scale where regression effects due to measurement error could be at play.

One of the most important implications of summer learning loss is its apparent relationship with poverty. With close to 500 citations in the literature, 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 math achievement over the summer, reading achievement patterns are economically moderated. Students from lower-income homes, on average, lose reading achievement over the summer, while students from higher-income homes stay the same or gain in reading achievement over the summer months. The authors estimate 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) find only a weak relationship between achievement change and socioeconomic status. Also, Bryk and Raudenbush (1988) find an opposite, negative relationship where summer losses in math 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 1996 meta-analysis, a major limitation of these results is that most of the recent studies that examine this relationship have been done using one, publically-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) use 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, it only focuses on the summer learning between kindergarten and first grade.

Most recently, Sandberg Patton and Reschly (2013) find that summer loss in grade two is moderated by family income and special education status. However, there is no evidence of systematic, disruptive summer loss in the other grade levels. The current research will not only contribute to the validity literature on Student Growth Percentiles, but also provide a fresh perspective on the relationship between summer learning and poverty.

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) finds that the impact the summer months 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) suggests summer learning loss is a potential factor contributing to the observed variability in the estimates across the two testing windows, this hypothesis is not formally tested.

A thorough search of the literature revealed only three studies that empirically explore effects of summer learning loss on estimates of teacher and school effectiveness. Downey, von Hippel, and Hughes (2008) introduce a new method of holding schools accountable called “impact,” which explicitly takes into account the variability in summer growth rates. These authors suggest that the difference between the school’s average summer growth rate (non-school rate) and the average in-school growth rate is a better indicator of school effectiveness than measuring student learning across the year. These authors find non-trivial differences between traditional 12-month growth models and their method which accounts for differences in summer learning.

McEachin and Atteberry (2014) use 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: 1) Is summer learning loss unevenly distributed across students and schools in a way that leads to systematic bias in aggregated measures of growth? 2) What is the relationship between growth estimates and student demographic variables? 3) How does the ranking of schools based on aggregate measures of growth change when the testing window moves from spring-to-spring to fall-to-spring? To address the first question, the authors modify the traditional spring-to-spring value-added model and a model similar to SGPs but derived using ordinary least squares estimation by changing the outcome variables to be the achievement gains/losses in the summer months. Significant school-effect coefficients in these models would indicate bias in the school estimates in typical VAMs and models similar to SGPs due to differential summer learning rates. Results showing that at least one school effect was significantly different from zero provided evidence for an uneven distribution of summer learning across schools. Results for the second research question show that across both models and content areas, correlations between percentage of free/reduced lunch students in the school and school effectiveness estimates decreased when calculated from fall-to-spring instead of spring-to-spring. Lastly, McEachin and Atteberry find that removing the summer months from the growth estimations increases the likelihood that schools with high percentages of free- and reduced-lunch eligible students are in the upper quintiles of effectiveness.

Gershenson and Hayes (under review) use the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K) to investigate the effect of summer months on value-added estimated classroom effects. The authors limit 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 show the Spearman correlations across the fall-to-spring and spring-to-spring value-added estimates are high, ranging between .8 and .9. The authors argue that even with high correlations such as those observed, the variability in teacher rankings across the estimation periods can result in misclassification of a non-trivial fraction of teachers, and support Hill (2009) in saying that these estimates are inappropriate for making high-stakes decisions. However, the authors find that including student characteristic variables and information related to student summer activities in the spring-to-spring value-added models does not improve the cross-period stability of the measured classroom effects. Rather than this finding being an indication that the model is sufficient for capturing summer learning patterns, it is more likely that the measured summer activity variables are not strongly related to summer loss.

The present study is comprehensive in its scope in order to build on the current literature and to help contribute to resolving conflicting findings. By analyzing data from two of the largest interim assessment programs in the country, the study builds in a cross-validation of any findings. Additionally, this study focuses on both content areas in all tested grade levels. Lastly, this study will serve to evaluate the validity of the SGP model for use in teacher evaluation systems rather than the more commonly studied value-added models.

Methods

Datasets

For this research project, datasets from two nationally distributed interim testing programs were acquired. Though no direct comparisons of quality are made across the two programs, as per the suggestion of the dissertation committee, the two data sources are hereto referred to as Test A and Test B. All personally identifiable information had been completely removed from both datasets before coming into the author's possession and both data sources 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 free- or reduced-lunch (FRL) program eligibility from the National Center for Educational Statistics (NCES) at the school level.

The Test A dataset contains a large sample from four states for the years 2009-2010 and was obtained through a company-sponsored data award. The Test B dataset, while smaller, represents sixteen states and was obtained through an internship with the National Center for the Improvement of Educational Assessment. The Test B sample was drawn from the company's complete dataset from the years 2012-2013. Subjects were selected for inclusion based on the following criteria: First, subjects must have had at least one measurement occasion recorded for each of the three testing windows, between May 1 and June 30, 2012, between August 1 and September 30, 2012, and between May 1 and June 30, 2013. When multiple measurement occasions occurred during the window, the last occasion was chosen in the spring 2012 and 2013 windows and the first occasion was chosen for the fall 2012 window. Secondly, subjects must have been in grades three through eight for spring 2013. Lastly, subjects must have been one of at least ten students who also met the other two criteria in their grade level at their respective schools. 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
Test A	2009	2009	2010	4	14905	716	14379	679	4-8
Test B	2012	2012	2013	16	10998	89	8880	95	3-8

Table 2

Distributions of Grade-Level Spring-to-Spring MGPs

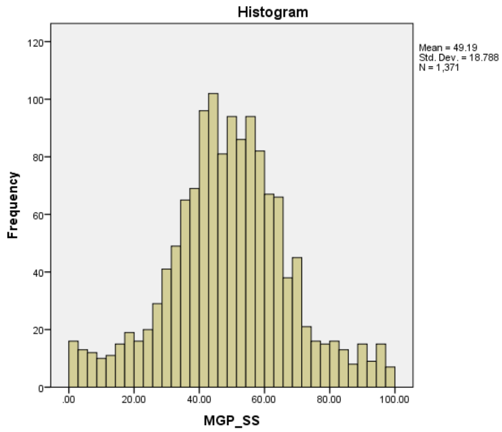
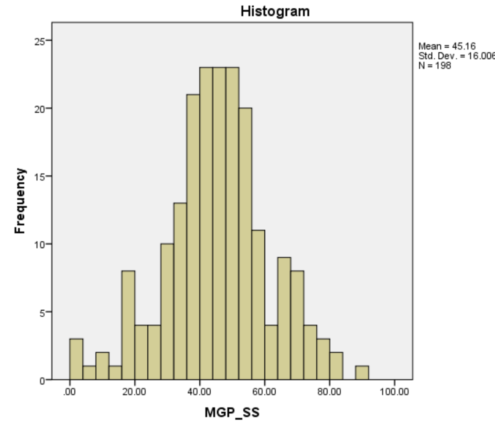
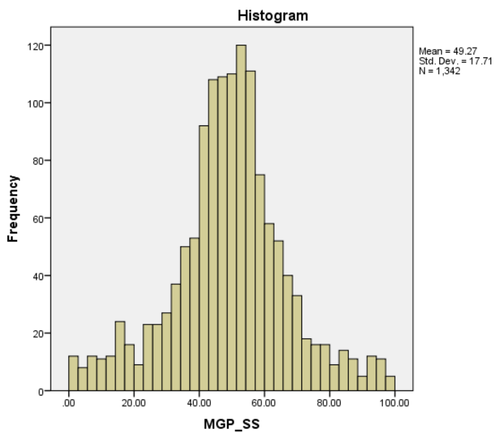
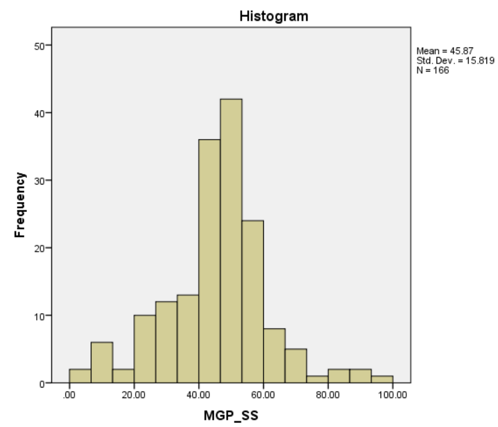
<table>
<tr><th></th><th>Test A</th><th>Test B</th></tr>
<tr><td>Math</td><td>
</td><td>
</td></tr>
<tr><td>ELA</td><td>
</td><td>
</td></tr>
</table>

Table 2 above shows the respective distributions of spring-to-spring mean Student Growth Percentiles for the grade-level units in each sample. While all four distributions are slightly leptokurtic, there does not seem to be a restriction of range problem in that the variability in the grade-level growth rates is sufficient for analysis and will support external validity.

Figure 2 below shows that the Test A sampling distribution for the percentage of students who are eligible for free- or reduced-lunch within the grade level closely follows the distribution of students nationally (NCES, 2014). While the Test B dataset is a bit more variable, this fluctuation is expected given the greater probability of sampling error in smaller datasets. Despite this deviation, the main purpose of examining the distribution is to ensure sufficient variability across the poverty continuum for the analyses. This condition is satisfied by the samples which will contribute to the internal validity of the statistical conclusions.

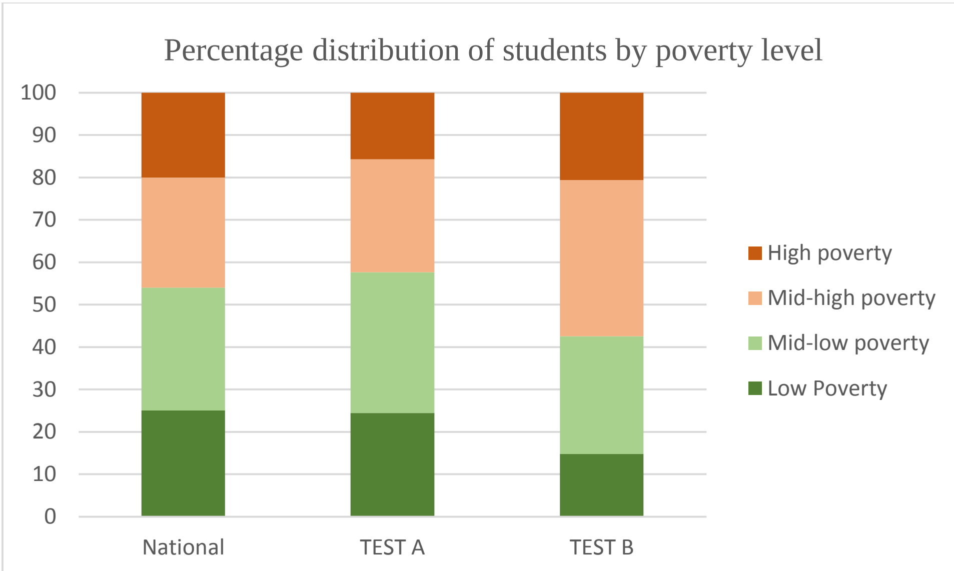


Figure 1. *Percentage Distribution of Students by Poverty Level. This bar chart compares national distribution of poverty levels to distribution in the datasets.*

Analyses

All of the analyses are completed separately for each dataset and each subject area resulting in four sets of results. The Test A and Test B datasets will serve to cross-validate any patterns identified within and across subject areas.

Research Question 1: Summer loss effect size. First, the effect size of summer loss is analyzed to describe the extent of summer loss present overall and by subgroups. Yen's (1986) statistic defined for purposes of comparing year-to-year growth along a vertical scale is used to estimate the effect size of summer loss from spring-to-fall. This statistic is calculated as:

$$\text{Effect Size} = \frac{\theta_{spring} - \theta_{fall}}{\sqrt{\frac{\sigma^2_{spring} + \sigma^2_{fall}}{2}}} \qquad (4)$$

where θ_{spring} and θ_{fall} are the vertical scale scores for the spring and fall achievement measurements, respectively, and σ^2_{spring} and σ^2_{fall} are the respective sample variances for the scores at each time point. Effect sizes are calculated at the student level for both content areas and each grade level for both samples. The effect size estimates serve as descriptors of actual loss; because of the standardization of the statistic, direct comparisons can be made. Additionally, factorial analyses of variance for both subjects and all grade levels separately identify any significant differences in effect size of summer loss across levels of prior achievement and poverty. Because free- or reduced-lunch information is only available at the school level for both datasets, these analysis are conducted with grade-level units of analysis. Only grade-level units where the number of students is greater than or equal to 10 are included in these analyses. Of those schools, three in each of the Test A datasets are missing free- or reduced-lunch eligibility information. For the ANOVAs, missing data is dealt with in a pairwise fashion. High-poverty grade-level units are defined using the National Center for Education Statistics standard: those with greater or equal to 75% of the students eligible for free- or reduced-lunch (Aud et al., 2010). Prior achievement is examined categorically by the average prior spring performance quartiles. These variables have been treated categorically in order to provide a descriptive summary of summer loss by subgroup. These analyses are done separately for both subjects and all grade levels to avoid violating the independence-of-observations assumption for the general linear model. An *a priori* alpha level for all analyses of $\alpha = .05$ is set. Follow-up pairwise comparisons are tested using Tukey's HSD as a type-1 error correction. Importantly, effect sizes of summer loss are compared to effect sizes of within-year growth using the same Yen's effect size statistic. These comparisons serve as validity checks that the instruments are sensitive to student learning and, additionally, can provide information of the potential impact of the summer loss.

Research Question 2: Predicting summer loss. While absolute loss estimates used in the analyses above serve as a first step in understanding the extent of summer loss within the datasets, growth percentiles calculated from spring-to-fall provide a normative description of student growth/loss over the summer months. Student Growth Percentiles are calculated for the spring-to-fall period to describe individual student movement in the distribution over the summer months. Low spring-to-fall SGPs indicate relative loss, while higher SGPs will indicate relative growth. For the Test B dataset, SGPs were calculated by a third-party vendor, the National Center for Improvement of Educational Assessment, using up to three prior scores and a national norm group. For Test A, SGPs were calculated by the author using one prior score and the sample as the norm group. The purpose of this analysis is to understand any systematic variance in the spring-to-fall MGPs due to poverty. In order to capture the most power with this analysis, all grade levels within school units are modeled simultaneously with a two-level hierarchical linear model. Note: all students with complete data for the analysis are included, school and grade level units are not limited to those where the number of students is greater or equal to 10.

$$SF\_SGP_{ij} = \beta_{0j} + \beta_{1j}*(\mathrm{G4})^{\dagger} + \beta_{2j}*(\mathrm{G5}) + \beta_{3j}*(\mathrm{G6}) + \beta_{4j}*(\mathrm{G7}) + \beta_{5j}*(\mathrm{G8}) + e_{ij} \qquad (5)$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ \beta_{2j} &= \gamma_{20} \\ \beta_{3j} &= \gamma_{30} \\ \beta_{4j} &= \gamma_{40} \\ \beta_{5j} &= \gamma_{50} \end{aligned}$$

$$SF\_SGP_{ij} = \beta_{0j} + \beta_{1j}*(\mathrm{G4})^{\dagger} + \beta_{2j}*(\mathrm{G5}) + \beta_{3j}*(\mathrm{G6}) + \beta_{4j}*(\mathrm{G7}) + \beta_{5j}*(\mathrm{G8}) + e_{ij} \qquad (6)$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}*(\%FRL_{j}) + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ \beta_{2j} &= \gamma_{20} \\ \beta_{3j} &= \gamma_{30} \\ \beta_{4j} &= \gamma_{40} \\ \beta_{5j} &= \gamma_{50} \end{aligned}$$

† This term is not included in the model for Test A

where SF_SGP_{ij} is the spring-to-fall Student Growth Percentile for student i in school j, β_{0j} is the mean SGP at school j in grade 3, and β_{1j} to β_{5j} are the effects for the respective dummy-coded grade levels with grade 3 and the reference group for Test B and grade 4 (since grade 3 information is not available) as the reference group for Test A. These variables are included to account for within-school variance and to avoid violating the independence-of-observations assumption. The remaining within-school variability in SGPs that cannot be explained by grade level is denoted with e_{ij}. This level-one equation models the within-school variability in SGPs, while the following six, level-two equations represent the between-school variability. For Model 5, the baseline model, there are no level-two predictors. In this case, the error term, u_{0j}, represents all between-school variability in mean SGPs (MGPs). In Model 6, the first level-two equation includes percent of students eligible for free- or reduced-lunch as a predictor for school-level MGPs: γ_{00} is the average MGP for students in grade three when zero percent of the students in the school are eligible for the federal free- or reduced-lunch program. The coefficient γ_{01} represents the change in γ_{00} for a school with 100 percent student eligibility for free- or reduced-lunch. A significant, negative coefficient would indicate a significant effect of poverty on summer learning loss. The error term for this equation, u_{0j}, represents the between-school variability in average MGP that cannot be accounted for by the percentage of students eligible for free- or reduced-lunch (%FRL). It is expected that u_{0j} from Model 6 will be smaller than u_{0j} from Model 5 if the percentage of students eligible for free- or reduced-lunch accounts for any between-school variability in MGPs. The remaining level-two coefficients, γ_{10} to γ_{50}, represent deviations from the grade 3 MGPs for the other grade levels, respectively. Significant coefficients indicate significant differences in the effect of poverty on MGPs due to grade level. The hierarchical linear model 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 math and English Language Arts, the hierarchical linear models are estimated separately for math and ELA for both the Test A and Test B datasets.

Research Question 3: Improving the implicit model. Lastly, it is hypothesized that by controlling for summer learning, the observed relationship between mean Student Growth Percentiles 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 (see pp. 3-7), and the hypothesized model would then be equation 2 from p. 4:

$$Y_i \perp Z_i \mid V_i \qquad (2)$$

where V_i is the spring-to-fall Mean Student Growth Percentiles that are related to both spring-to-spring MGPs, Y_i, and the percentage of students in school eligible for free- or reduced-lunch, Z_i. The resulting orthogonal, or more likely diminished, relationship between Y_i and Z_i is represented by "⊥". This model is tested with grade levels as the unit of analysis and also at the school level for both subjects, and testing programs using a series of bivariate and partial correlations:

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

Relationship 1 is the squared correlation between spring-to-spring MGPs and the percentage of students eligible for free- or reduced-lunch and relationship 2 is the same squared correlation but controlling for spring-to-fall MGPs. The differences between the bivariate and partial squared correlations are tested for statistical significance using an *F* test, with an *a priori* type-1 error rate of $\alpha = .05$. Because there will be a unique hypothesis associated with each test, a type-1 error correction is not necessary.

Results

Research Question 1

What is the effect size of summer learning loss? How do average summer losses compare by subject, grade-level, prior achievement, and poverty-level?

First, the effect size of summer loss is analyzed to describe the extent of summer loss present overall and also disaggregated by subgroups. Yen's (1986) statistic defined for purposes of comparing year-to-year growth along a vertical scale is used to estimate the effect size of summer loss from spring-to-fall. This statistic is calculated as:

$$\text{Effect Size} = \frac{\theta_{fall} - \theta_{spring}}{\sqrt{\frac{\sigma^2_{fall} + \sigma^2_{spring}}{2}}} \tag{4}$$

where θ_{fall} and θ_{spring} are the vertical scale scores for the fall and spring achievement measurements, respectively, and σ^2_{fall} and σ^2_{spring} are the respective variances for the scores at each time point. Therefore, summer loss effect size is reported in standard deviation units and can be interpreted similarly to a Cohen's *d* effect size. When summer loss is present, the effect size statistic is negative in value. Table 3 below shows the effect sizes of summer loss broken down by testing program, grade level, and subject area with the unit of analysis as the student. Because this is simply a descriptive analysis, it is more informative to examine the amount of loss experienced at the individual student level, rather than averages across groups of students. Note: The effect sizes are calculated for the summer prior to the listed grade level (e.g., estimates listed for Grade 3 represents loss from spring of Grade 2 to fall of Grade 3).

Table 3

Effect Size of Summer Loss

Grade	Test A Math (n = 14905)	Test B Math (n = 10998)	Test A ELA (n = 14379)	Test B ELA (n= 8880)
3		-0.100		0.088
4	-0.077	-0.245	-0.053	0.049
5	-0.094	-0.163	-0.028	0.034
6	-0.115	-0.195	-0.049	0.026
7	-0.035	-0.137	-0.005	0.046
8	-0.011	-0.039	-0.005	0.051

In general, though these effect sizes appear small, they are not, on average, smaller than what has been reported in the previous literature. Summer losses are larger in math than in ELA, and more apparent in Test B than Test A. The differences in effect sizes across the two tests is notable. More substantial loss is found for math in the Test B dataset, while gains, rather than losses, are shown for Test B ELA.

To help interpret the magnitude of the summer loss, these effect sizes are compared to the effect sizes of within-school year growth in Figures 2 and 3 on the next page.

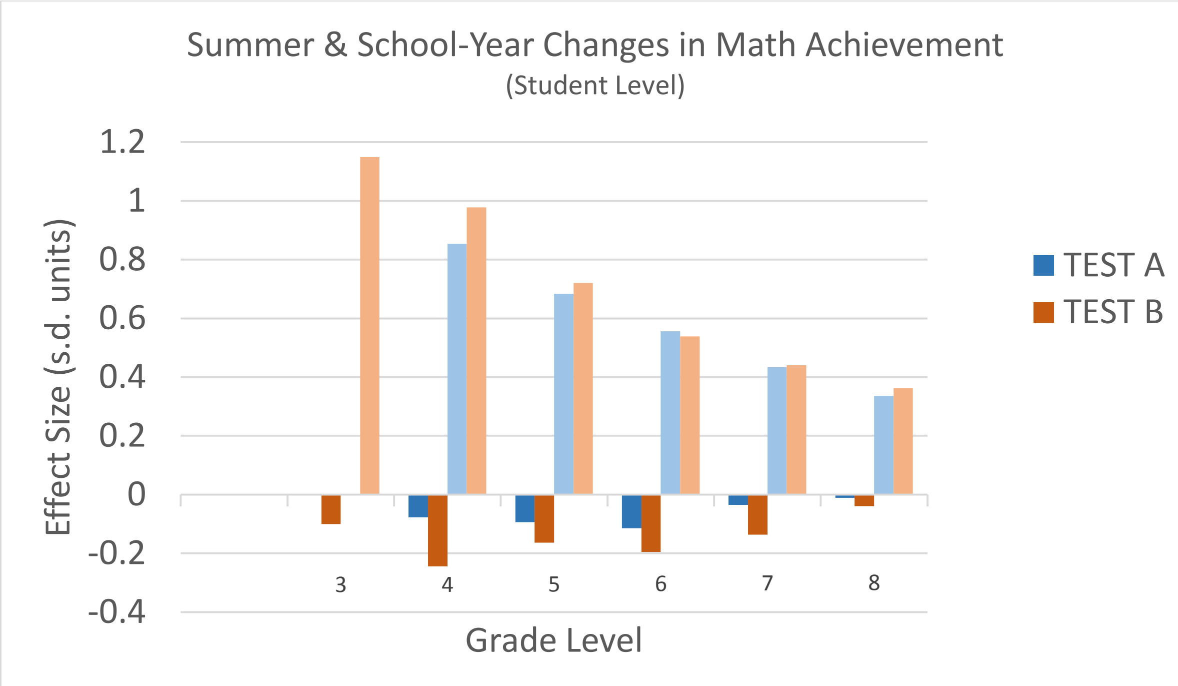


Figure 2. Summer and School-Year Changes in Math Achievement

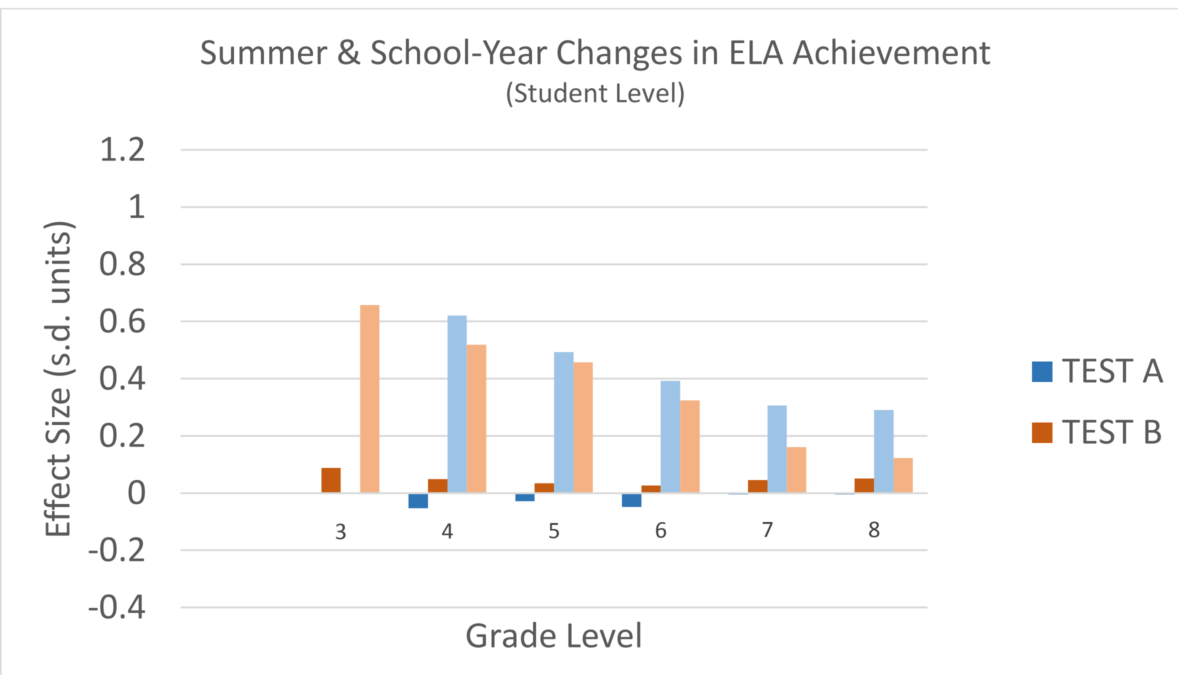


Figure 3. Summer and School-Year Changes in ELA Achievement

Figures 2 and 3 reveal that, on average, as grade level increases, the effect sizes for both summer loss and within-year growth decrease. The effect size for summer loss ranges between 3.18% and

36.31% of the within-school-year growth for math. For ELA, effect size of summer achievement change ranges from gains of 41.57% to losses of 12.43% of school-year growth. The complete table of these percentages can be found in Table 14 in Appendix A.

To analyze effect sizes defined by poverty level and prior achievement, a series of analyses of variance are run. Because no student-level information for free- or reduced-lunch eligibility is available, the unit of analysis is grade units within schools. Aggregated grade units were limited to those with a minimum sample size of 10 students. To avoid violating the assumption of independence of observations, grades 3-8 are analyzed separately. A hierarchical analysis, capitalizing on the power of the nested data, was not run because the goal of the current research question is descriptive in nature rather than inferential. The dependent variable is summer loss effect size while the independent variables are poverty with two levels and prior achievement with four levels. High-poverty grade levels are defined using the National Center for Education Statistics standard: those with greater or equal to 75% of the students eligible for free- or reduced-lunch (Aud et al., 2010). Prior achievement, which is measured by the prior spring score, is examined categorically by the average prior spring performance quartiles. Tables 4-7 below show the marginal means of summer loss effect size by poverty level and prior achievement quartiles. Additionally, the adjusted R-squared estimates for the ANOVAs are included. Adjusted R-squared estimates are reported to enhance the comparability of the results due to the discrepancy in sample sizes across the testing programs. Significant results of follow-up, pairwise comparisons with Tukey's HSD type-1 error correction are indicated with asterisks. Note: The effect sizes are calculated for the summer prior to the listed grade level (e.g., estimates listed for Grade 4 represents loss from spring of Grade 3 to fall of Grade 4).

Table 4

Test A Math - Marginal Effect Sizes by FRL and Quartile

	Grade	4 (n=376)	5^+ (n=359)	6 (n =236)	7 (n=210)	8 (n=190)
FRL	0	-.0783	-.0693*	-.1386	-.0587	.0412**
	1	-.0619	-.0918	-.1189	.0658	-.0592
Quartile	1	.0544*	.1195**	-.0383	.0665	.2260
	2	-.0293	-.0974	-.1342	-.0049	.0648*
	3	-.1660	-.1343	-.1483	-.0609	-.0481
	4	-.1601	-.1783	-.2224	-.1788	-.1318
R^2 (adj.)		.020	.124	.000	.011	.184

*significantly different from the following categories, $\alpha < .05$
**significantly different from the following categories, $\alpha < .01$
+significant interaction effect

Grades 5 and 8 provide evidence that groups with high poverty showed significantly greater losses than groups with typical or low poverty. This is noteworthy because, on average, there is an increase in loss as prior performance increases; prior spring performance negatively correlates with effect size of loss, ($r_{grade\ 5} = -.256$, $p < .01$; $r_{grade\ 8} = -.378$, $p < .01$) where loss is indicated by a negative effect size. However, there is also negative correlation between FRL and performance, ($r_{grade\ 5} = -.423$, $p < .01$; $r_{grade\ 8} = -.295$, $p < .01$). The combination of these negative correlations is what drives, in grade 5, a significant interaction effect between performance quartile and the poverty variable. The high-poverty schools, across the achievement continuum, are losing achievement over the summer, and this loss is significantly more than for the low-poverty schools.

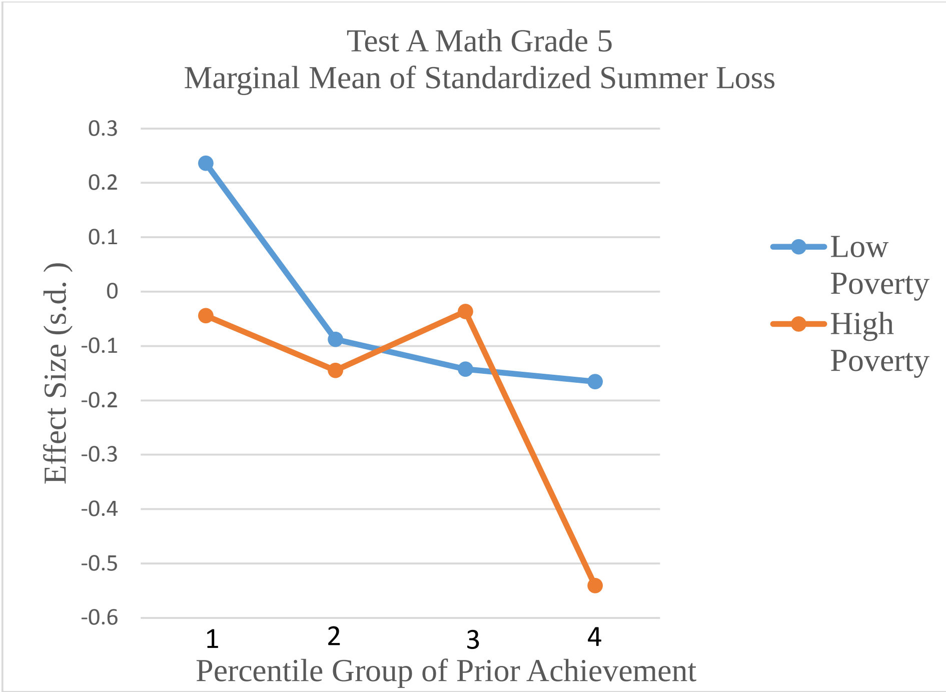


Figure 4. Significant Interaction Effect between FRL and Prior Achievement

Table 5

Test B Math - Marginal Effect Sizes by FRL and Quartile

	Grade	3 (n=46)	4 (n=52)	5 (n=40)	6 (n=34)
FRL	0	-.0507	-.2377	-.1407	-.5259
	1	.0051	-.0986	-.2373	-.0249
Quartile	1	.3163*	.2940*	-.0862	-.2816
	2	-.1131	-.3440	-.0932	-.1112
	3	-.1423	-.2192	-.2564	-.8109
	4	-.1837	-.5318	-.2238	-.3526
R^2 (adj.)		.190	.239	.000	.000

*significantly different from the following categories, $\alpha < .05$

Due to sample sizes fewer than 30, the ANOVAs were not run for grades 7 and 8 in the Test B Math dataset. For this dataset, with the exception of the lowest performers in grades 3 and 4, who saw significant gains, FRL and prior achievement do not significantly predict summer loss.

Table 6

Test A ELA - Marginal Effect Sizes by FRL and Quartile

	Grade	4 (n=364)	5 (n=355)	6 (n=236)	7 (n=201)	8 (n=186)
FRL	0	-.0845	-.0535	-.0573	.0008	-.0133
	1	-.0519	.0899	-.0289	-.0215	-.0282
Quartile	1	.0438[4]	.1362**	-.0064	.3216**	.2047[44]
	2	-.0842	-.0920	-.1040	-.2015	.0763
	3	-.0961	-.0614	-.0550	-.0071	-.0926
	4	-.1757	-.0916	-.0467	-.1212	-.2504
R^2		.027	.035	.000	.111	.042

**significantly different from the following categories, $\alpha < .01$

[4]significantly different from the fourth category, $\alpha < .05$

[44]significantly different from the fourth category, $\alpha < .01$

The only significant finding from the Test A ELA analysis is that the lowest quartile of achievers experienced summer growth, which in all grades but 6, was statistically significant. In grade 4 and 8, the first quartile of achievers experienced gains, which was significantly different from the losses shown in the fourth quartile. In grades 5 and 7, the first quartile gains were significantly different from the losses shown in all of the following quartiles. There are no significant differences in summer loss between groups with high poverty and those with typical or low poverty.

Table 7

Test B ELA - Marginal Effect Sizes by FRL and Quartile

	Grade	3 (n=34)	4 (n=43)	5 (n=39)
FRL	0	.1608	.0991*	.0682
	1	.0805	-.0526	.1873
Quartile	1	.0222	.1081	.2119
	2	.2484	.1390	-.1242
	3	.1235	.1349	.1761
	4	.1826	.1088	.0949
R^2 (adj.)		.000	.129	.000

*significantly different from the next category, $\alpha < .05$

Again, due to limited sample sizes in the Test B dataset for grades 6-8, the analyses of variance for these grades are omitted. While prior achievement quartile in grade 4 was significant overall for predicting summer loss, no follow-up pairwise comparisons were significant. In grade 4, grade units with high poverty showed significant summer losses in ELA compared to their low-poverty counterparts with marginal summer gains.

To summarize the findings for research question 1, in general, effect sizes for summer loss in both datasets are small but on the same order of magnitude as what has been reported in previous literature. Losses were larger for mathematics, and for Test B ELA, summer gains were detected. Additionally, the effect sizes for Test B are larger in magnitude than those found for Test A. The series of analyses of variance revealed that neither free- or reduced-lunch nor prior achievement are strong predictors of summer loss. The adjusted R-squared estimates range from 0 to .239. However, where significant effects were identified, FRL status was associated with greater loss. The effect of poverty may be somewhat suppressed due to the unexpected, positive relationship between prior achievement and summer loss.

Research Question 2

What proportion of variance in summer learning patterns can be accounted for by poverty?

While absolute loss estimates used in the analyses for research question 1 serve as a first step in understanding the extent of marginal summer loss within the datasets, growth percentiles calculated from spring-to-fall provide a normative description of student growth/loss over the summer months. Student Growth Percentiles were calculated for the spring-to-fall period and describe individual student movement in the distribution over the summer months. For the Test B dataset, SGPs were calculated by a third-party vendor, the National Center for Improvement of Educational Assessment, using up to three prior scores and a national norm group. For Test A, SGPs were calculated by the author using one prior score and the sample as the norm group. Low spring-to-fall SGPs indicate relative loss, while higher SGPs indicate relative growth. The purpose of the second set of analyses is to understand any systematic variance in the spring-to-fall MGPs due to poverty. In order to capture the most power with these analyses, all students within each testing program dataset are modeled simultaneously with a series of two-level hierarchical linear models shown below:

$$SF\_SGP_{ij} = \beta_{0j} + \beta_{1j}*(\mathrm{G4})^{\ddagger} + \beta_{2j}*(\mathrm{G5}) + \beta_{3j}*(\mathrm{G6}) + \beta_{4j}*(\mathrm{G7}) + \beta_{5j}*(\mathrm{G8}) + e_{ij} \quad (5)$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$
$$\beta_{5j} = \gamma_{50}$$

$$SF\_SGP_{ij} = \beta_{0j} + \beta_{1j}*(\mathrm{G4})^{\ddagger} + \beta_{2j}*(\mathrm{G5}) + \beta_{3j}*(\mathrm{G6}) + \beta_{4j}*(\mathrm{G7}) + \beta_{5j}*(\mathrm{G8}) + e_{ij} \quad (6)$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(\%FRL_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$
$$\beta_{2j} = \gamma_{20}$$
$$\beta_{3j} = \gamma_{30}$$
$$\beta_{4j} = \gamma_{40}$$
$$\beta_{5j} = \gamma_{50}$$

where SF_SGP_{ij} is the spring-to-fall mean Student Growth Percentile for student i in school j, β_{0j} is the mean SGP at school j in grade 3, and β_{1j} to β_{5j} are the effects for the respective dummy-coded grade levels with grade 3 as the reference group for Test B and grade 4 (since grade 3 information is not available) as the reference group for Test A. These variables are included to account for more within-school variance, and to avoid violating the independence-of-observations assumption. The remaining within-school variability in SGPs that cannot be explained by grade level is denoted with e_{ij}. This level-one equation models the within-school variability in SGPs, while the following six, level-two equations represent the between-school variability. For Model

‡ This term is not included in the model for Test A

5, the baseline model, there are no level-two predictors. In this case the error term, u_{0j}, represents all between-school variability in mean SGPs (MGPs). In Model 6, the first level-two equation includes percent of students eligible for free- or reduced-lunch as a predictor for school-level MGPs, γ_{00} is the average MGP for students in grade three when zero percent of the students in the school are eligible for the federal free- or reduced-lunch program. The coefficient γ_{01} represents the change in γ_{00} for a school with 100 percent student eligibility for free- or reduced-lunch. A significant, negative coefficient would indicate a significant effect of poverty on summer learning loss. The error term for this equation, u_{0j}, represents the between-school variability in average MGP that cannot be accounted for by the percentage of students eligible for free- or reduced-lunch (%FRL). It is expected that u_{0j} from model 6 will be smaller than u_{0j} from model 5 if the percentage of students eligible for free- or reduced-lunch accounts for any between-school variability in MGPs. The remaining level-two coefficients, γ_{10} to γ_{50}, represent deviations from the grade 3 MGPs for the other grade levels, respectively. Significant coefficients indicate significant differences in the effect of poverty on MGPs due to grade level.

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 math and English Language Arts, the hierarchical linear models are estimated separately for math and ELA for both the Test A and Test B datasets. Tables 8-10 show the results of the four hierarchical models that were run using the software, HLM 7.01 (Raudenbush, Bryk & Congdon, 2013). Tables 8 and 9 report the coefficients for the remainder of the level-two equations for Model 6 and their associated significance levels for Test A and Test B, respectively.

Table 8

Grade-Level Coefficients for Model 6, Test A

	Fixed Effect	Est.	Std. error	*t*-ratio	*df*	*p* value
Test A Math	G5, γ_{10}	0.375	0.972	0.386	14049	0.7
	G6, γ_{20}	1.472	1.227	1.199	14049	0.23
	G7, γ_{30}	2.432	1.38	1.763	14049	0.078
	G8, γ_{40}	3.284	1.673	1.963	14049	0.05
Test A ELA	G5, γ_{10}	-0.079	0.733	-0.107	13666	0.915
	G6, γ_{20}	-0.02	0.898	-0.022	13666	0.982
	G7, γ_{30}	0.268	0.911	0.294	13666	0.769
	G8, γ_{40}	0.261	0.957	0.273	13666	0.785

Table 9

Grade-Level Coefficients for Model 6, Test B

	Fixed Effect	Est.	Std. error	*t*-ratio	*df*	*p* value
Test B Math	G4, γ_{10}	-1.042	2.404	-0.433	10904	0.665
	G5, γ_{20}	-3.441	2.422	-1.42	10904	0.156
	G6, γ_{30}	-2.179	2.472	-0.882	10904	0.378
	G7, γ_{40}	-5.858	2.602	-2.251	10904	0.024
	G8, γ_{50}	-4.737	2.597	-1.824	10904	0.068
Test B ELA	G4, γ_{10}	1.762	3.001	0.587	8192	0.557
	G5, γ_{20}	-0.325	3.549	-0.092	8192	0.927
	G6, γ_{30}	-3.085	3.818	-0.808	8192	0.419
	G7, γ_{40}	-2.846	5.412	-0.526	8192	0.599
	G8, γ_{50}	2.458	4.128	0.596	8192	0.551

Grade level does not have a large effect on within-school average SPGs. There are no significant grade-level coefficients in the ELA datasets. In the math datasets, grades 7 and 8 in dataset A and grade 7 in dataset B have significant coefficients. However, the direction is opposite, for Test A, average spring-to-fall MGPs are significantly larger in grades 7 and 8 for schools with no students eligible for free- or reduced-lunch, while the grade 7 coefficient for Test B is negative. In general, because grade level does not have a large, meaningful effect on average SGPs, interpreting γ_{01} for school-wide effects is appropriate.

Table 10 reports the coefficient estimate for the %FRL variable, its significance, and the proportion of variance it can account for in between-school variability in MGPs. Note: Sample sizes are marginally smaller here than in the effect size calculations for research question 1 because three schools in each of the Test A datasets are missing information about free- or reduced-lunch eligibility, and a five schools in the Test B ELA dataset are missing information about SF_SGPs. Missing data analysis and interpretation can be found in Appendix B.

Table 10

%FRL Estimates for Models 5 and 6

	n level-1	n level-2	γ_{01}	t ratio^{+}	p-value (one-tailed)	$u_{0j(Baseline)}$	$u_{0j(FRL)}$	%level-2 variance accounted for by FRL
Test A Math	14766	713	-7.997	-4.666	<0.001	52.862	50.048	5.32%
Test A ELA	14343	676	-5.684	-3.959	<0.001	20.480	18.253	10.88%
Test B Math	10998	89	-3.067	-3.368	0.183	40.358	39.99	0.91%
Test B 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 four FRL coefficients are significant. The significant coefficients range from -5.684 to -7.997, and can be interpreted as the number of points lost in the spring-to-fall MGP scores for schools with 100% FRL eligibility compared to schools with no students eligible for free- or reduced-lunch. For Test A math, the percentage of students eligible for free of reduced lunch accounts for 5.32% of the variability in between-school spring-to-fall MGPs, while for the Test A and B ELA samples, this term accounts for 10.88% and 2.69% of the variability, respectively. This means that though the magnitude of summer loss in ELA is generally less, as seen in research question 1, the relationship with FRL may be equal or stronger than in math. 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.

Research Question 3

Does controlling for loss over the summer months reduce the magnitude of the relationship between mean Student Growth Percentiles (MGPs) and student-level poverty?

It was hypothesized that if mean Student Growth Percentiles are biased estimates of teacher or school effectiveness, then controlling for summer learning would reduce the observed relationship between mean Student Growth Percentiles and student poverty. This hypothesis stems directly from the forward causal inferences discussed in the introduction (see pp. 3-7), and the hypothesized model would then be equation 2 from p. 4:

$$Y_i \perp Z_i \mid V_i \tag{2}$$

where V_i is the spring-to-fall Mean Student Growth Percentiles that are related to both spring-to-spring MGPs, Y_i, and the percentage of students in the school eligible for free- or reduced-lunch, Z_i. The resulting orthogonal, or more likely diminished, relationship between Y_i and Z_i is represented by "⊥". This model is tested at both the grade and school levels separately for both subjects and testing programs. Since aggregated MGPs are being examined in these analyses, only units with 10 or more subjects are included in the analysis. The hypothesis is tested with a series of 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 percentage 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 α = .05. The *F* test statistic is calculated in the following way:

$$F = \frac{(r_{YZ}^2 - r_{YZ.V}^2)/1}{(1 - r_{YZ}^2)/(n-2)} \quad (7)$$

This ratio represents the percentage of residual variance from the bivariate correlation that can be accounted for by controlling for poverty, and since it is a ratio of two variances, with an expected value of 1, an *F* distribution is assumed. 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):

$$F = \frac{(R_{Y.AB}^2 - R_{YA}^2)/k_B}{(1 - R_{Y.AB}^2)/(n - k_A - k_B - 1)} \quad (8)$$

In equation 7, 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.

Tables 11 and 12 show the grade level results for Test A and Test B respectively. Grades 7 and 8 in math and grades 6-8 in ELA were not analyzed in the Test B dataset due to small sample sizes. Table 13 shows the results for all four datasets when the unit of analysis is the school. Note: sample sizes are slightly smaller for this analysis than the factorial analyses of variance because three of the schools in each of the Test A datasets are missing information about FRL eligibility.

Table 11

Results for Test A Datasets with Grade-Level Units of Analysis

		n	$r_{Z,\text{performance}}$	$r_{Y,Z}$	$r^2_{Y,Z}$	$r^2_{Y,Z.V}$	F	p-value
Math	grade 4	376	-0.438	-0.194	0.038	0.009	11.358	0.000
	grade 5	359	-0.423	-0.157	0.025	0.006	6.860	0.005
	grade 6	234	-0.352	-0.035	0.001	0.000	0.271	0.668
	grade 7	207	-0.378	-0.097	0.009	0.004	1.209	0.198
	grade 8	188	-0.295	-0.115	0.013	0.008	1.057	0.228
ELA	grade 4	364	-0.460	-0.174	0.030	0.027	1.307	0.181
	grade 5	355	-0.433	-0.080	0.006	0.005	0.334	0.583
	grade 6	235	-0.313	-0.122	0.015	0.005	2.287	0.084
	grade 7	199	-0.270	-0.066	0.004	0.002	0.458	0.468
	grade 8	183	-0.372	-0.050	0.003	0.002	0.087	1.293

Table 12

Results for Test B Datasets with Grade-Level Units of Analysis

		n	$r_{Z,\text{performance}}$	$r_{Y,Z}$	$r^2_{Y,Z}$	$r^2_{Y,Z.V}$	F	p-value
Math	grade 3	46	0.042	-0.313	0.098	0.091	0.351	0.56
	grade 4	52	-0.309	-0.181	0.033	0.035	-0.133	
	grade 5	40	-0.08	0.087	0.008	0.024	-0.633	
	grade 6	34	-0.479	0.125	0.016	0.011	0.155	0.929
	grade 3	46	0.042	-0.313	0.098	0.091	0.351	0.56
ELA	grade 3	34	-0.242	-0.102	0.010	0.002	0.284	0.642
	grade 4	43	-0.400	-0.064	0.004	0.002	0.083	1.319
	grade 5	39	-0.223	-0.009	0.000	0.001	-0.049	

Table 13

Results with School-Level Units of Analysis

	n	$r_{Z,\text{performance}}$	$r_{Y,Z}$	$r^2_{Y,Z}$	$r^2_{Y,Z.V}$	F	p-value
Test A MATH	387	-0.420	-0.250	0.062	0.008	22.177	0.000
Test A ELA	382	-0.475	-0.272	0.074	0.026	19.500	0.000
Test B MATH	67	-0.205	-0.093	0.009	0.002	0.434	0.484
Test B ELA	70	-0.291	-0.092	0.009	0.021	-0.859	

As shown in Tables 11 and 12, significant differences between the bivariate and partial squared correlations occur at grades 4 and 5 in math for the Test A datasets. This means that after controlling for spring-to-fall MGPs, the correlation between grade-level spring-to-spring MGPs and student poverty 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 grade level in the math Test A dataset. In the rest of the grade levels for both subjects, controlling for the summer months did not significantly reduce the correlation between spring-to-spring MGPs and poverty. In three cases, Test B grades 4 and 5 math 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 Test A math dataset, the observed correlation between spring-to-spring MGPs and poverty is not due to systematic differences in learning patterns over the summer months.

In Table 13, where school-level units are analyzed, the Test A datasets for both math and ELA show significant reductions in MGP bias after controlling for summer loss. Results being somewhat stronger for the school-level analysis is expected as previous research has shown, and this study confirms, that the correlation between MGPs and FRL generally increases as the aggregate unit of analysis increases (Marland, 2014). In general, across the three tables of results, the Test B datasets do not produce any significant findings. This does not seem to be a result of purely a lack of power to detect an effect because the effect sizes themselves are smaller for the Test B dataset. As a data check procedure, the correlations for FRL and performance are shown in the results tables. With the exception of grades 3 and 5 in the Test B Math dataset, the correlations are on the same order of magnitude for both testing programs. The low correlations between FRL and math achievement in grades 3 and 5 for Test B call into question the validity of the results for those grade levels across all analyses. The sample sizes for these two grade levels in the math dataset for Test B are small, which means that the low correlations could be a result of sampling error. The correlations between achievement and FRL for the other grade levels and datasets are only slightly below what would be expected based on the effect sizes reported for aggregated data in a 2005 meta-analysis that found correlations ranging from .11 to .85 with a mean of .60 with a standard deviation of .22 (Sirin, 2005). This may indicate that the effects of poverty may be somewhat suppressed in the datasets due to either inaccuracies in data collection, or, perhaps a lack of generalizability of these data to the greater population.

Discussion

The main issue that 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 is 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, nationally-distributed commercial interim testing programs, this hypothesis was investigated by answering a series of three research questions:

1. What is the effect size of summer learning loss? How do average summer losses compare by subject, grade-level, prior achievement, and poverty-level?
2. What proportion of variance in summer learning patterns can be accounted for by poverty?
3. Does controlling for loss over the summer months reduce the magnitude of the relationship between mean Student Growth Percentiles (MGPs) and student-level poverty?

Results from the first research question reveal that though, in general, summer loss effect sizes would be considered small, ranging from -.245 to .088, they represent a substantial portion of within school-year progress. For example, the summer gains in grade 8 for the Test B ELA dataset represent almost 42% of typical reading growth within the school year. This means that though effect sizes may seem small, they have real potential to greatly influence annual growth estimates. Additionally, there were substantial differences across the outcome measurements. Greater amounts of summer loss was detected in the Test B Math dataset, and also, greater effect sizes, but in the positive direction for the Test B ELA dataset. Because the variability in effect sizes across datasets is bi-directional, the difference is more likely due to differences in the sensitivity of the achievement scales for detecting growth, rather than differences in student populations. One factor that may influence test sensitivity is alignment to the instructional curriculum. A test will only detect summer losses/growth if it accurately measures the construct in question. This is only one possible explanation for why cross-test variability is observed, and future research on that collects content-related validity evidence for these two tests should be explored.

While the overall magnitudes of summer loss are important to investigate, the central hypothesis of this research focuses on comparative summer loss. In general, for both math and ELA, the effect sizes of summer loss and within-year growth decrease as the grade-level increases. Summer learning loss seems to be more of an issue at the lower grade levels. Additionally, across both datasets, summer losses are greater in mathematics than English Language Arts. For the Test A dataset losses in ELA are essentially zero, while the Test B dataset shows summer growth in ELA. Gains in summer reading are not unprecedented and also found by a recent 2008 study by Helf, Konrad, and Algozzine. These authors attribute the reading summer gains to the recent boost in programming offered to address summer learning loss. Summer setback awareness has only spread since 2008 and it has even made it onto First Lady Michelle Obama’s agenda. In 2014, Michelle Obama launched a “*Let’s Read. Let’s Move*.” campaign which is specifically designed to “combat summer reading loss and childhood obesity” (Corporation for National and Community

Service, 2015). It is possible that with all the focus on preventing summer learning loss, especially for those students in typically underserved communities, the programming is working. This emphasis on serving students in low-performing districts may also contribute to explaining why we see relative gains in reading and math for the students at the lowest end of the achievement scale.

Previous literature suggests that summer loss is moderated by prior achievement and poverty. In order to descriptively understand how summer loss may be different for student groups, a series of factorial analyses of variance were run. In general, both poverty and prior achievement are poor predictors of absolute summer loss with mostly very small R^2 estimates for the analyses. On average, prior achievement quartile and NCES poverty status only accounted for approximately 7% of the variance in summer learning loss. In eight of the seventeen analyses of variance, prior achievement was found to be a predictor of summer loss, however, this occurred in an unanticipated direction. In all cases where prior achievement was a significant predictor for summer loss, the relationship was positive. This means that the students at the bottom end of the achievement scale were more likely to gain achievement over the summer months than their higher achieving peers. This finding is not unprecedented (see Klibanoff & Haggart, 1981) and may be an artifact of both increased summer programming for low-achieving students, and regression to the mean effects. Regression to the mean will occur if many students are scoring at the highest or lowest obtainable scale score; more generally, regression effects can be a result of measurement error. Further research into the relationship between achievement and summer learning loss is warranted.

Research has shown and the current data support that achievement is negatively correlated with indicators of poverty. Despite the positive relationship between prior achievement and summer loss, the analyses of variance provide evidence for a negative relationship between poverty and summer loss. All of the significant mean differences in summer loss effect sizes between grade-level units that are and are not classified as high-poverty are in the hypothesized direction. For the majority of the analyses of variance run, poverty cannot predict differences in summer loss effect sizes, but for those three cases where poverty is a significant factor (Test A Math grades 5 and 8, and Test B ELA grade 4), grade-level units classified as high-poverty show significantly greater losses than grade-level units without at least 75% of students eligible for free- or reduced-lunch. These significant differences are troubling because they indicate systematic change in summer learning patterns for units classified as high-poverty. In the Test A Math data, a significant interaction effect was found for the FRL and prior achievement quartile variable. The most important take away from this type of interaction, and the variability of results in general across the set of ANOVAs run for this research, is that summer loss effect size depends on a combination of characteristics of students within the unit. The analyses for research question 1 are important for generally understanding the size of summer loss and how it varies across student groups, but are not the most powerful way to understand these differences. Instead of looking at observed loss, the analyses for the second research question attempted to quantify the percentage of variance in summer loss that can be attributed to differences in poverty by analyzing normative loss.

Spring-to-fall student growth percentiles were calculated and analyzed to better understand normative summer loss using a series of hierarchical linear models. The results showed that school-level poverty was a significant predictor of mean student growth percentiles (MGPs) for both math and ELA in the Test A dataset and for ELA in the Test B dataset. This means that school-level differences in MGPs are systematically influenced by the level of poverty of the students within the school. There was not a significant relationship between poverty and spring-to-fall MGPs for math in the Test B dataset. In the significant analyses, the percentage of between-school variability in MGPs accounted for by the FRL variable ranged from 2.69% in Test B ELA, to 5.32% in Test A math, to 10.88% in Test A ELA. This means that though poverty can explain some between-school differences, the majority of the differences in summer learning patterns are left unexplained. While informative, this analysis cannot detect how much influence these explained portions of variance may have on annual estimates of MGPs. Research question 3 investigates the influence of the summer months on the correlation between spring-to-spring MGPs and the percentage 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 run and their differences tested. The correlations between school-level MGPs and percentage of students eligible for free- or reduced-lunch are generally greater than those observed for grade-level units. This confirms earlier research that suggests the amount of bias in the MGP estimate increases as the unit of aggregation gets larger (i.e., further from the classroom). The results follow logically that partialing out the summer months from the bivariate correlations had a larger effect at the school level than when analyzed for the grade-level units. At the grade level, of all the analyses, only grades 4 and 5 for Test A math showed a significant reduction in the correlations between spring-to-spring MGPs and FRL. This means that when grade levels are the units of analysis, bias in MGPs due to summer learning loss may only be a real issue in math and at the lower grade levels. At the school level, both Math and ELA for Test A showed a significant reduction in the squared correlation between MGPs and FRL once spring-to-fall MGPs had been controlled for, moving from values of r^2 = .062 to r^2 = .008 for math and from r^2 = .074 to r^2 = .026 for ELA. Though the correlations between spring-to-spring MGPs are small in magnitude to start with (r =-.250 and r = -.272 for math and ELA respectively) the degree of reduction in the correlation represents the 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 school evaluation, when calculated the way they were for Test A, will likely result in meaningful misspecification of school quality with a downwards bias for those schools with higher percentages of students eligible for free- or reduced-lunch. This means that due in part of the variability over the summer months, schools that serve more disadvantaged students will receive lower MGPs that are not completely reflective of their quality but instead, in part due to factors outside of their realm of control.

Interestingly, while it is not unexpected given the results of the previous analyses that the Test B dataset showed no significant reduction in squared correlation, the results indicate that the significance is not likely due to lack of power. Instead, the MGPs calculated for Test B seem to be truly less influenced by poverty, as the amount of shared variance between spring-to-spring MGPs and FRL to start with is only .009 for both math and ELA. Because the correlations between prior achievement and FRL are of similar magnitudes across both tests, this lack of significance does not seems to be only the result of a data issue with the FRL variable. Instead, this may be an artifact of the way the SGPs were calculated and normed, using multiple prior achievement measures and a national dataset, rather than conditioning on a single prior score and using within-sample norming, as was done for Test A.

Due to a lack of reliable links between students and teachers, this study analyzes variance in MGPs at the grade level and school level. While this limits generalizability to some extent, it is not likely that any systematic variance in grade-level MGPs would not also occur at the classroom level, to a somewhat lesser extent. The findings of this study serve as a framework for understanding the effects of summer learning loss on MGPs and warrant future 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 yes, but the degree to which this is a real issue varies across testing programs, grade levels, subject areas, and unit of analysis. Even with seemingly low marginal effect sizes for summer learning loss, annual estimations of normative student growth can still be substantially impacted. Correlating spring-to-spring mean Student Growth Percentiles with summer loss reveals significant relationships ranging from r = -.310 to r = -.662 across the four datasets (see Appendix A for the complete results). These correlations can be interpreted to mean that summer learning loss does influence annual estimates of student growth, and, because the summer loss that was detected does not seem to be primarily a function of student-level poverty, simply controlling for student-level poverty will not likely alleviate the issue much. Poverty was only able to explain between zero to 11% of the between-school variability in summer learning patterns as measured by spring-to-fall MGPs. Based on strong correlations between annual MGPs and summer loss, and the variability in findings for research questions 1 and 2, the first policy recommendation is that when designing estimates of student growth to be used for teacher or school 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 VAMs) than more directly controlling for summer loss itself. The results of this study showed that poverty is a significant factor, but only one factor that can explain the influence of the summer months on MGPs. Instead of controlling for student level poverty in the model, a more effective way of reducing bias in the model introduced by summer learning loss would be to control for the summer months. A natural way to control for the summer months would be to implement a fall-to-spring accountability testing program. Future research should be conducted to investigate whether growth estimates calculated for the academic school year, rather than based on annual measurements, would lead to more valid estimates for teacher evaluation.

Secondly, based on the findings from research question 3, both the number of prior observations and the size and generalizability of the norm group likely matter when calculating SGPs. For the Test B 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 using more than one prior achievement score and a national norming group, as was used when calculating the SGPs for Test B, may be able to better account for the variability in summer learning patterns. If a student who loses over the summer months one year is more likely to lose the next year, than using multiple years of data and a large norm group may accurately norm the student's scores to make better determinations of normative student growth. This conclusion is certainly not groundbreaking and has already been shown using simulated data (see Castellano & Ho, 2013), but it does re-emphasize the real-world importance of maintaining large, longitudinal datasets in order to improve the accuracy of Student Growth Percentiles and growth models in general.

Limitations

The first major limitation of this study is the reliance on vertical scales to quantify change in academic achievement. Though the Student Growth Percentile model does not need or rely on a vertical scale, it does assume that student achievement is measured accurately. Inherent problems arise when building a unidimensional scale to measure achievement and growth over multiple years. Because a unidimensional scale is the result of projecting all dimensions onto a single line, the resultant scaled score can be considered a weighted composite of knowledge on each of the dimensions (Wang, 1986). As the nature of the construct changes across years, extending the linear scale without accounting for the multidimensionality of student learning will inevitably result in a loss, likely substantial, of information about student achievement (Reckase, 2004). While this remains a significant limitation of the studied data, unidimensional vertical scales are commonly used for accountability purposes and thus worthy of study.

Secondly, both tests used for this study are administered under low-stakes conditions. This presents issues related to standardization of implementation and generalizability. While the administration and results of these tests are designed to mimic the state summative test, fidelity of implementation cannot be assured. As a result, these tests might be, in general, less reliable for measuring student achievement than the state tests with stricter enforcement of standardization procedures. Additionally, because there are no stakes for students, teachers, and administrators associated with the results of the test, student effort may be less than optimal. Both of these problems present validity issues when trying to generalize the study results to the tests used for state accountability purposes. The MAP and STAR datasets are chosen due to their wide use and fall testing programs; however, future research should be conducted with the actual instruments that are used to calculate measures of student growth for teacher evaluation purposes. Both the Partnership of Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) have plans for implementing assessment programs that measure student achievement across the school year. These forthcoming programs will provide rich datasets that should be used for further investigation of the relationship between summer learning patterns and aggregate measures of student growth.

Lastly, due to a lack of reliable links between students and teachers, this study analyzes variance in MGPs at the grade and school levels. While this limits generalizability to some extent, it is not likely that any systematic variance in grade- and school-level MGPs would not also occur at the classroom level. Because of the likelihood of increased sampling error at the classroom level, any patterns observed at the grade level have the potential to be exaggerated or unobserved in any individual classroom. The findings of this study serve as a framework for understanding the effects of summer learning loss on MGPs and will likely warrant future research at the classroom level.

References

Allinder, R. M., & Eicher, D. D. (1994). Bouncing back: Regression and recoupment among students with mild disabilities following summer break. *Special Services in the Schools*, *8*(2), 129-142.

Aud, S., Hussar, W., Planty, M., Snyder, T., Bianco, K., Fox, M. A., ... & Drake, L. (2010). The Condition of Education 2010. NCES 2010-028. *National Center for Education Statistics*.

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. W. (2009). SGP: Student growth percentile and percentile growth projection/-trajectory functions [Computer software manual]. (R package version 0.0-4).

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.

Borman, G. D., & Dowling, N. M. (2006). Longitudinal achievement effects of multiyear summer school: Evidence from the Teach Baltimore randomized field trial. *Educational Evaluation and Policy Analysis*, *28*(1), 25-48.

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.

Briggs, D., & Betebenner, D. (2009, April). *Is growth in student achievement scale dependent?* Presented at the annual meetings of the National Council on Measurement in Education, San Diego, CA.

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.

Camera, L. (2014, August 21). *Waiver states can seek a year's delay in using tests for teacher evaluations.* Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2014/08/ed_announcment.html

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.

Corporation for National and Community Service (2015, January 12). *Let's Read. Let's Move.* Retrieved from http://www.serve.gov/?q=site-page/lrlm.

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/

Elder, H. E. (1927). The effect of the summer vacation on silent-reading ability in the intermediate grades. *The Elementary School Journal*, *27*, 541-546.

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.

Gabriel, T. (2010, September 2). A celebratory road trip by education secretary, *New York Times*, p. A24.

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. (under review). The implications of summer learning loss for value-added estimates of teacher effectiveness.

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*.

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher*, *39*(2), 120-131.

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.

Koenker, R. (2005). *Quantile Regression*. New York, NY: Cambridge University Press.

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.

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.

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 Development Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, *29*(1), 117-120.

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.

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, August 29). *ESEA Flexibility: Guidance for renewal process for window 1 and 2 states.* Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/flex-renewal/index.html

Walberg, H. J., & Tsai, S. L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*(3), 359-373.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional item response data. In *The Office of Naval Research Contractors Meeting, Gartlingburg*.

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

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299-325.

Appendix A: Additional Tables of Results

Table 14

Change in Summer Months as Percentage of Within-Year Growth

Grade	TEST A Math	TEST B Math	TEST A ELA	TEST B ELA
3		-8.74%		13.46%
4	-9.07%	-25.06%	-8.49%	9.37%
5	-13.71%	-22.68%	-5.72%	7.49%
6	-20.63%	-36.31%	-12.43%	8.01%
7	-8.00%	-31.06%	-1.57%	28.37%
8	-3.18%	-10.84%	-1.80%	41.57%

Table 15

Correlations between Annual Grade-Level MGPs and Summer Loss (more loss is more positive)

Grade level	n	Test A Math	n	Test B Math	n	Test A ELA	n	Test B ELA
3			34	-.590**			46	-.172
4	376	-.470**	43	-.118	364	-.423**	52	.231
5	359	-.450**	39	-.570**	355	-.455**	40	-.487**
6	236	-.425**	23	-.595**	236	-.453**	34	-.475**
7	210	-.363**	13	-.512	201	-.498**	14	-.429
8	190	-.310**	14	-.208	186	-.461**	12	-.662*

*statistically significant at the $\alpha = .05$ level

**statistically significant at the $\alpha = .01$ level

Table 16

Correlations between Annual School-Level MGPs and Summer Loss (more loss is more positive)

Test	n	r
Test A Math	388	-.407**
Test B Math	67	-.337**
Test A ELA	383	-.438**
Test B ELA	70	-.177

**statistically significant at the $\alpha = .01$ level

Appendix B: Missing Data Analysis

Three schools in each of the two Test A datasets are missing information related to free- or reduced-lunch eligibility. The FRL-related parameter estimates for these schools, and their associated students and grade-level units, are calculated based on pairwise deletion for the first set of analysis, and listwise deletion for the second two sets of analyses. Tables 17 and 18 below explore the appropriateness of this missing data treatment by comparing the average achievement of students in the affected schools to all other students in the datasets.

Table 17

Mean Math Achievement Comparisons for Test A FRL Missing Data

		n	Mean difference	*t* (adjusted for unequal σ^2)	*df* (adjusted for unequal σ^2)	*p value*
Spring09	Not missing Missing	14766 139	13.06	11.15	142.36	<.001
Fall09	Not missing Missing	14766 139	12.83	10.47	142.01	<.001
Spring10	Not missing Missing	14766 139	10.83	8.45	141.60	<.001

Table 18

Mean ELA Achievement Comparisons for Test A FRL Missing Data

		n	Mean difference	*t*	*df*	*p value*
Spring09	Not missing Missing	14343 36	9.85	3.74	14377	<.001
Fall09	Not missing Missing	14343 36	8.21	3.02	14377	0.003
Spring10	Not missing Missing	14343 36	8.87	3.56	14377	<.001

In both the math and ELA datasets, independent samples *t* tests reveal significant differences in achievement between students at schools with and without free- or reduced-lunch eligibility information. In both cases, the students without FRL information score significant higher in all three testing windows than the rest of the samples. Because FRL is known to correlate with achievement, this test shows that our data is missing not at random. This finding should be taken into account when interpreting the analyses for all three research questions in this dissertation. However, due to the small proportion of missing data, power is not compromised. This is especially true because this data is not likely to have come from high-poverty schools, where subgroup sample sizes are smaller.

In the Test B ELA dataset, 593 students are missing spring-to-fall student growth percentiles. To test differences between students with missing data and those without, mean spring-to-spring and fall-to-spring growth percentiles were compared across the two groups. The results of this analysis are below in Table 19.

Table 19
Mean SGP Comparisons for Test B ELA SF_SGP Missing Data

		n	Mean difference	*t*	*df*	*p value*
SS_SGP	Not missing	14343	-5.42	-4.64	8878	<.001
	Missing	36				
FS_SGP	Not missing	14343	-5.95	-4.94	8878	<.001
	Missing	36				

The results of the two, independent samples *t* tests show that students who are missing spring-to-fall Student Growth Percentiles have significantly lower SGPs for the other two calculation periods. Because Student Growth Percentiles are highly correlated across estimation windows, our data for this variable are not missing at random. This may introduce bias into the parameter estimates related to research questions 2 and 3. However, because our missing data is such a small proportion of the total sample, power is not likely reduced, and bias due to listwise deletion is likely small.

Appendix C: IRB Approval

KU THE UNIVERSITY OF KANSAS
Research

APPROVAL OF PROTOCOL

December 18, 2014

Susan Gillmor
scgillmor@ku.edu

Dear Susan Gillmor:

On 12/18/2014, the IRB reviewed the following submission:

Type of Review:	Initial Study
Title of Study:	Effect of the Summer Learning Loss on Aggregate Estimates of Student Growth
Investigator:	Susan Gillmor
IRB ID:	STUDY00001976
Funding:	None
Grant ID:	None
Documents Reviewed:	• Gillmor_Initual Submission,

The IRB approved the study on 12/18/2014.

1. Notify HSCL about any new investigators not named in the original application. Note that new investigators must take the online tutorial at https://rgs.drupal.ku.edu/human_subjects_compliance_training.
2. Any injury to a subject because of the research procedure must be reported immediately.
3. When signed consent documents are required, the primary investigator must retain the signed consent documents for at least three years past completion of the research activity.

Continuing review is not required for this project, however you are required to report any significant changes to the protocol prior to altering the project.

Please note university data security and handling requirements for your project:
https://documents.ku.edu/policies/IT/DataClassificationandHandlingProceduresGuide.htm

You must use the final, watermarked version of the consent form, available under the "Documents" tab in eCompliance.

Sincerely,

Stephanie Dyson Elms, MPA
IRB Administrator, KU Lawrence Campus

Appendix D: Memorandum of Understanding with NWEA

MEMORANDUM OF UNDERSTANDING BETWEEN
NORTHWEST EVALUATION ASSOCIATION
AND
Susan C. Gillmor, University of Kansas

THIS MEMORANDUM OF UNDERSTANDING ("MOU"), is entered into this 8th day of April, 2013 by and between **Northwest Evaluation Association** ("NWEA"), located at 121 NW Everett Street, Portland, Oregon and **Susan C. Gillmor** ("Data Award Recipient") located at Department of Psychology and Research in Education. School of Education, J.R. Pearson hall, 1122 West Campus Road, University of Kansas, Lawrence, KS (hereinafter also referred to individually as Party and collectively as Parties).

Recitals

NWEA is an educational 501(c)(3) not-for-profit corporation that provides assessment tools and test development services to school districts and engages in ongoing, supportive relationships with partnering school districts and education agencies throughout the United States.

Susan Gillmor is a recipient of the Kingsbury Center Data Award.

Agreement

In consideration of the mutual commitments herein contained, the Parties agree as follows:

1. The Data Award Recipient shall protect and undertake to keep confidential all information received from NWEA in accordance with the Growth Research Database (GRD) Confidentiality Agreement. The Data Award Recipient and her advisor **John Poggio, PhD** shall sign the GRD Confidentiality Agreement prior to receiving any data from NWEA and shall follow the data handling procedures therein, a copy of which is attached hereto and made a part herewith (Attachment 1).

2. The Data Award Recipient and NWEA shall develop a project proposal and data specification based on Data Award Recipient's proposal attached as Attachment 2 for the use of the data from the GRD. The GRD data fields provided to Data

Award Recipient will not contain school or student identifiable data. The access, storage and use of the data must not exceed two years from the delivery date of the data from NWEA. The Data Award Recipient shall provide periodic status updates to NWEA. Any extension beyond the initial two year expiration date must be pursuant to a new agreement.

3. The Data Award Recipient shall send a copy of any and all reports or papers that are produced using GRD data and NWEA may publish the report(s) on its web site.

4. This MOU is governed by the laws of the State of Oregon, without regard to that state's conflict of laws provisions.

5. This MOU commences on the date hereof and continues in effect until it is superseded by a formal contract or until it is terminated by mutual agreement of the Parties, or in two years, whichever date occurs first. In any event, the confidentiality obligations set forth in this MOU and in the Confidentiality Agreement attached hereto are binding and survive termination of this MOU.

IN WITNESS WHEREOF, the Parties have caused this Memorandum of Understanding to be executed by their duly authorized officers as of the day and date first above written.

Northwest Evaluation Association

Jeffrey P. Strickler — Susan C. Gillmer

Vice President of Corporate Services and Chief Operating Officer — Graduate Student/Ph.D Candidate

Date: 4/30/2014 — Date: 04/29/2014

Updated 4/28/2014