EdWorkingPaper No. 19-82

School's Out: The Role of Summers in Understanding Achievement Disparities

Allison Atteberry
University of Colorado Boulder

Andrew McEachin
RAND Corporation

The field is generally aware of summer learning loss (SLL)—that student learning slows during the summer. Yet surprisingly little consensus exists on basic questions about SLL, especially given recent concerns about measurement artifacts in foundational SLL studies. Nearly all prior SLL work examines how summers contribute to racial/ethnic or socioeconomic gaps. However, these factors only account for about 4 percent of the variance in summer learning rates. We use a unique dataset with millions of students across eight grades to document the full spread of SLL and examine how differential SLL contributes to where students end up in the 8th grade achievement distribution. We find dramatic variability in SLL, with decrements accruing to the same students over time.

VERSION: May 2020

School's Out: The Role of Summers in Understanding Achievement Disparities

Allison Atteberry, PhD*
CU Boulder School of Education
249 UCB
Boulder, CO 80309
allison.atteberry@colorado.edu

Andrew McEachin, PhD
RAND Corporation
1776 Main Street
Santa Monica, CA 90407
mceachin@rand.org

Note updated citation:
Atteberry, A., & McEachin, A. (in press, 2020). School's out: The role of summers in understanding achievement disparities. *American Educational Research Journal*.

May 2020

ALLISON ATTEBERRY, PhD, is an assistant professor of research and evaluation methodology at the University of Colorado–Boulder School of Education. Her work addresses persistent patterns of inequality in key educational pivot points, including early childhood education, access to effective teaching, and summer learning loss. Contact: allison.atteberry@colorado.edu.

ANDREW McEACHIN, PhD, is a policy researcher in the Economics, Statistics, and Sociology Department at the RAND Corporation and a professor at the Pardee RAND Graduate School. His research focuses on the determinants of persistent achievement gaps, as well as evaluating the effect of popular responses by policymakers and educators to reduce these gaps.

The project was supported in part by the Kingsbury Data Award funded by the Kingsbury Center at the NWEA, as well as the Smith Richardson Foundation. We are particularly grateful for Dr. Nate Jensen, Research Scientist at the Kingsbury Center, for his help in acquiring the MAP assessment data. The research was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B100009 to the University of Virginia. All errors are solely attributable to the authors.

*=Corresponding author

Children experience vastly different home environments prior to formal schooling (Gilkerson & Richards, 2009; Kaushal, Magnuson, & Waldfogel, 2011; Kornrich & Furstenberg, 2013) and thus arrive to kindergarten with a wide range of starting skills (Lee & Burkam, 2002; Magnuson, Meyers, Ruhm, & Waldfogel, 2004). Yet even once school begins, children *continue* to spend a significant portion of their school-age years outside the school setting. That out-of-school time is concentrated in the summer months—a time when schools play little to no direct role in children's lives. Instead, children return to full-time care of their families, with vastly different options and preferences for how children spend this time (Gershenson, 2013). Student achievement disparities[1] may grow dramatically during the summer, when child experiences appear most diverse.

We use a novel dataset with over 200 million test scores for students across the U.S. to explore whether the "fanning out" of achievement from grade 1 to 8 occurs while students are in school or during the intervening summers. The field is generally aware of the phenomenon called summer learning loss (SLL)—that student learning slows during the summer. Less apparent, however, is how little consensus actually exists on basic questions about SLL. Moreover, many of the canonical findings on SLL have recently been called into question based on measurement concerns that apply to data used in most prior SLL research (von Hippel & Hamrock, 2019).

At a time when even fundamental questions in the SLL literature need to be revisited, our analyses also contribute a unique focus on the *total variability* in SLL—a surprisingly understudied phenomenon. Nearly all prior SLL work focuses on how summers contribute to race/ethnic or socioeconomic (SES) gaps.[2] However, these factors together only account for about 4% of the variance in summer learning rates (von Hippel et al., 2018). These gaps deserve our attention,[3] yet a sole focus on these gaps alone misses important big-picture questions about the SLL landscape. Herein, we zoom out to explore the full spread of SLL experiences and examine how differential SLL contributes to where students end up in the achievement distribution at the end of 8th grade.

Even before concerns arose about possible measurement artifacts in SLL, surprisingly few aspects of SLL have been well-established. For instance: Do students, on average, actually *lose* ground during the summer, or just exhibit no gain (i.e., flat)? What proportion of a student's school-year gain tends to be lost in the summer that immediately follows? Is the magnitude of SLL similar across students, or do some students exhibit gains while others actually lose ground? Does this vary by grade level? Do summer losses accrue to the same students year after year? We tackle these questions using a set of achievement scores that are less susceptible to the measurement concerns raised by von Hippel and Hamrock (2019). These foundational questions have theoretical implications for the production of outcome inequality, as well as practical implications for where researchers and policy makers look for opportunities to disrupt this stratification process.

We focus on estimating the *total variability* in SLL across students, relative to school-year gains. Describing this total (or unconditional) variance is important for at least four reasons: (1) Summers will only contribute to widening achievement disparities if students exhibit meaningful *variation* around the typical summer pattern. We find that SLL does vary dramatically across students. (2) Because of this wide variability, mean SLL patterns—those that most researchers, policymakers, and practitioners are familiar with—do not characterize most students' summer experiences very well. (3) We find evidence that the same students are likely to lose ground from summer to summer, suggesting a non-random accumulation of summer decrements. (4) Prior work finds that even a full vector of student demographics, home characteristics, prior achievement, and a list of summer activities account for only 13% of the variation in SLL (Burkam et al., 2004).[4] In other words, SLL appears to vary greatly, but race and class—which have been the main focus of prior SLL research—are an important but limited part of the story.

Contribution of Current Study

Data provided by the Northwest Evaluation Association (NWEA) allow us to estimate means and variances in SLL across 8 grade levels, using a dataset with over 200 million test scores for nearly

18 million students in 7,500 districts across all 50 states in a very recent time period (2008 through 2016). NWEA's Measures of Academic Progress (MAP) scores are IRT-based, computer adaptive in all grades, and cover a broader range of content than scores used in prior SLL research. The use of MAP scores, in and of itself, represents a timely contribution to the field of SLL, because von Hippel and Hamrock (2019) have recently shown that newer data sources and scaling practices can dampen and sometimes even *reverse* some of the long-standing inferences about SLL gaps. They also argue that the above features of NWEA's test scores can make achievement gain inferences less susceptible to measurement artifacts. Their work has raised troubling questions about the robustness of what we thought we knew about SLL. The current study is among a new wave of SLL research to revisit our foundational knowledge about SLL, and our findings reaffirm the existence and importance of this phenomenon.

We use this powerful dataset in a hierarchical student growth modeling framework to characterize the contribution of SLL to end-of-school achievement disparities. Specifically, we answer the following four questions:

(1) *On average, how do learning gains during the school-year compare to gains/losses during the summer across grade levels?*
(2) *Of more relevance to the current investigation, how much do students vary in terms of how much they gain or lose?*
(3) *Do the same students tend to exhibit summer learning loss year after year, or are these gains/losses randomly distributed?*
(4) *How large is the role of summers in producing end-of-school outcome disparities?*

With respect to the questions posed above, we *do* find that some students maintain their school-year learning rate throughout the summer, while others can lose almost as much ground as they had gained in the preceding school-year. We show that even if all the inequality in school-year learning rates could be entirely eliminated, students would still end up with very different achievement levels due to SLL alone. Our findings also suggest that negative summer decrements tend to accumulate to the same students over time: We find that more than twice as many students exhibit 5 years of consecutive summer *losses* (as opposed to no change or gains) than one would expect if summer losses were independently distributed across students and grades. Furthermore, these consecutive losses add up to a sizeable impact on where students end up in the achievement distribution: In a five year period, the average student in this group ultimately loses nearly 40% of their total school-year gains during the intervening summers.

In what follows, we first (I) situate the contributions of the current study within existing SLL literature. (II) Next, we introduce this unique dataset and how it compares to the broader U.S. public school population. We also describe a significant primary data collection activity undertaken to address a methodological concern in SLL research about the dates on which tests are taken (more on this below). (III) In the Methods section, we present our multilevel model and key parameters. (IV) The Results section is organized by the four research questions previously described. (V) The Conclusion provides a reflection on our results relative to prior SLL findings, the study limitations, and implications for future research.

I. Evidence on SLL

There are logistical challenges to studying SLL: The data provided by annual end-of-school-year statewide testing systems, which are most often used by researchers, lack the fall datapoint needed to separate learning gains between the school-year and the summer. Opportunities to investigate SLL have necessarily been limited to idiosyncratic samples (e.g., one city), specific years, or particular grades (e.g., only after grade K/1). Figure 1 provides an overview of the data used across 17 key SLL studies, including if each one focuses on seasonal patterns in White-Black achievement gaps, SES gaps, and/or unconditional variance in achievement—the latter of which is our focus and is relatively unique. Figure 1 also highlights some advantageous features of the current dataset in terms of size, number of grades included, and recency.

Much has been written about SLL (see e.g., Gershenson (2013) for a particularly thorough recent overview; or Cooper, Nye, Charlton, Lindsay, and Greathouse (1996) for a meta-analysis of early studies). Today, there is a common understanding among policy-makers, researchers, and practitioners that, during the summer, students lose some knowledge and skills acquired during the school-year.[5] The seminal SLL research comes from two key studies: Heyns' study of the summer after 5th grade for about 3,000 students in 42 Atlanta schools from 1970 to 1972 (Heyns, 1978), and Entwisle and Alexander's study of the summers after grades 1 - 4 for about 750 students in 20 Baltimore schools from 1982 to 1987 (Alexander et al., 2001; Alexander et al., 2007; Entwisle & Alexander, 1992). These studies documented that, on average, students learning rates slow during the summer. Heyns found that average 5th and 6th grade school-year gains in Atlanta were positive (about 60% of a national norm for one year of achievement gains), while summer after 5th grade gains were either flat or very modestly negative, depending on cohort. Alexander et al. (2001) used a multilevel, quadratic individual growth curve model to document slower summer (versus school-year) learning. The authors have continued to follow their Baltimore sample through adulthood and have found that early differences in summer learning are predictive of later life outcomes such as high school completion and college-going (Alexander et al., 2007). The findings from these studies became the definitive word on summer setback, raising awareness of the phenomenon and the role it plays in growing educational inequality.[6]

More recently, researchers have used the Early Childhood Longitudinal Study Kindergarten Class (ECLS-K) 1998-99 or 2010-11 cohorts to study SLL (Benson & Borman, 2010; Burkam et al., 2004; Downey et al., 2004; Downey, von Hippel, & Hughes, 2008; Quinn, 2014; Quinn et al., 2016; Quinn & Le, 2018; von Hippel & Hamrock, 2019; von Hippel et al., 2018). The advantage of the ECLS-K is that the samples are nationally-representative (which NWEA is not). This constituted a major step forward for the SLL literature. The ECLS-K data have a few limitations: While the current study includes grades 1 to 8, ECLS-K only covers the summer after K or 1st grade, which limits one's view of how SLL accumulates as students move through school. In addition, because of the sampling methods used for ECLS-K (e.g., on average, only 3.2 students per K school have SLL estimates[7]), clustered analyses seeking to estimate the variability in SLL are not straightforward. The current NWEA data is therefore a useful complement to the ECLS-K data, since the weaknesses of each one is a strength of the other.

One of these ECLS-K studies—by von Hippel et al. (2018)—has a unique analysis that is particularly relevant to the current study. These authors also examine the unconditional variance in SLL at the student level (most like the current study) through the summer after grade 1. Interestingly, they find that the variation in achievement *shrinks* over that time. They also find that the variation in achievement arises more in summers than in school-years. The current study extends these analyses through grade 8, and we consider how results from the two compare.

Another recent study by von Hippel and Hamrock (2019), which compares SLL racial and SES gap findings[8] across three datasets, warrants more detailed discussion. This paper has raised some important questions about SLL, since the authors show that measurement artifacts can lead to quite different conclusions about how much gaps grow over time. For instance, when they use two different scalings[9] of math achievement scores available in ECLS-K 1998-99, one indicates that student-SES gaps grow by 83% between grade 1 through 8 while their preferred scaling suggests these gaps *decrease* by 27%. When von Hippel and Hamrock (2019) conduct that same analysis using BSS achievement scores—which the authors posit have several undesirable measurement properties—they find that SES gaps appear to grow 369%. The question of whether SES gaps grow more in summers versus school-years, however, appears to be less sensitive to variations in data sources and scalings: In most permutations, they confirm the finding from the original BSS data that SES gaps grow faster in the summer period versus the school year.

The von Hippel and Hamrock (2019) study is also particularly relevant to our current analysis, because they use a subsample from NWEA's Growth Research Database (GRD). The full NWEA data that we use may not necessarily be comparable to the GRD subsample; the GRD is much smaller (e.g., 25 school districts versus 7,500) and has a shorter panel (2 versus 8 years). For the current analysis, the key point from their study is the authors' argument that the features of the NWEA/GRD data make achievement gain inferences less susceptible to measurement artifacts (e.g., IRT scaling, computer adaptive in all grades, broader content). Their exploration of how measurement properties affect the study of SLL would bolster confidence in our results.

Finally, though both of these papers use similar data, they focus on different questions: Whereas the current study describes the degree of total variation in SLL, von Hippel and Hamrock (2019) focus in on racial and SES gaps (although due to data limitations one cannot look at student-level SES gaps with the GRD data). As mentioned above, race and SES appear to play an important, but quite small part in explaining variability in SLL.

We are aware of one other peer-reviewed study that uses a subsample of NWEA data to explore SLL. Rambo-Hernandez and McCoach (2015) juxtapose the school-year and summer growth trajectories of initially high- and average-achieving students.[10] Their results suggest that high-achieving students exhibit steadier growth throughout the panel, while "average-achieving" students actually grow faster during the school year but lose more during summers.

In sum, the extant research on SLL took an important leap forward in the late twentieth century, and it now seems to be experiencing a resurgence of interest, particularly spurred by the availability of the ECLS-K data. This new work improves upon the methods used in prior work (e.g., by taking into account test timing, considering measurement artifacts), updates the evidence to the a more recent period, and covers a nationally representative sample (in grades K and 1).

The current paper continues in this tradition, building off the various methodological advances in this domain. First, NWEA's Measures of Academic Progress (MAP) tests are designed to be vertically-scaled assessments of math and reading achievement, which facilitates an examination of student growth across grades (Quinn, 2014; von Hippel & Hamrock, 2019). In addition, we undertake a massive primary data collection effort to recover over 44,000 district-year calendar dates for the start and end of the school-year, allowing us to make crucial adjustments to SLL estimates on a large scale. We also implement a set of multi-level models that we think connect more clearly to the central research questions in this domain: The coefficients ("fixed effects" in the language of HLM) correspond to school-year gains and summer losses, while the variance components allow us to characterize a plausible range of gain/losses one should expect across students during those periods. These variance components connect directly to our primary research question: The larger the variation in summer losses across students, relative to the school-year gains, the more summers are the time when end-of-school achievement disparities arise.

Figure 1 compares key aspects of the current study to prior work. The defining feature of the current study is our unique focus on documenting the scope and seasonality of the total (unconditional) variation in achievement across U.S. students. The current dataset also provides data on over 18 million students across a wider range of grades than possible in prior work. In addition, the NWEA dataset comes from the 2008 through 2016's post-accountability era—a time in which it is at least conceivable that the dynamics of access to quality schooling have changed.

II. Data and Sample

NWEA Data

The current study primarily uses data from the NWEA's MAP assessment. The dataset contains math and reading scores based on a computer adaptive test designed to serve as part of a formative, benchmarking data system, used in about 32,000 schools located in 7500 districts across all 50 states in the U.S. The MAP assessment is used as a supplementary tool to aid schools' in improving their instruction and meeting students' needs, not as the high-stakes test of record. Because the MAP assessment is intended to monitor students' progress throughout the school-year, it is administered in both the fall and the spring.[11]

NWEA's MAP test is designed so that its scores can be expressed on a vertical scale (which the NWEA calls the RIT scale), with the intent that the scale can be used to support equal-interval interpretations. In theory, the vertical scale allows comparisons of student learning across grades and over time, while the equal-interval property of the scale ensures that a unit increase in a student's score represents the same learning gain across the entire distribution. It is worth noting that there are many different ways of designing and calibrating a vertical scale, and there is little consensus with regard to the best methods for evaluating the properties of the scale (Briggs, 2013; Briggs & Dadey, 2015; Briggs & Domingue, 2013; Briggs & Weeks, 2009). Therefore, our findings regarding changes across grades assume NWEA's vertical scale is valid. However, much of the paper concerns itself with comparing learning gains in the same grade (that is, a given school-year relative to the subsequent summer).

The full dataset used for the current study comes from 7,685 U.S school districts that administered the MAP assessment during the nine years between 2008 and 2016. Different districts administer MAP in different grades, however the NWEA full dataset includes 203,234,153 test scores for 17,955,222 students who took a test between grades K and 11. The dataset includes students' race, gender, math and reading MAP scores, number of items attempted and correctly answered, duration of the test, grade of enrollment, and the date of test administration. The file does not include indicators for whether the student is an English Language Learner, belongs to the federal Free- and Reduced-Price Lunch program, or receives special education services. For this reason, the current dataset is not well-suited to study achievement gaps along these dimensions.

Adjustments to NWEA RIT Scores

Students do not take MAP tests *exactly* on the first and last day of school, but rather typically 3 to 6 weeks before/after the school-year starts/ends. As a result, some of the time between the spring and fall administrations of the test—what one would mislabel as summer time—is actually spent in school. While the NWEA dataset does include the test date, crucially, it does not include school-year start or end dates.

We therefore conducted a large-scale data collection effort to record the start- and end-date in every district in a subset of 11 states with the greatest use of MAP assessments. We found 23,223 school-year start dates and 20,807 school-year end dates—about 77% of the district-year calendar dates in those 11 states from 2008 to 2015. In later years, NWEA also began to collect school-year start and end dates. Together, these efforts allowed us to collect actual calendar start/end dates for 50.3% of the observed school-years for the entire NWEA dataset. Based on that data, we also extrapolate likely dates for other districts.[12] Following practices in prior SLL studies, we then use these calendar data to make a linear projection of each students' score on the first and last day of the school-year. For more information about this process, including a description of our approach to collecting this data, the percent of actual dates recovered, our extrapolation process, our score projection process, and how study results differ when using observed scores instead of projected scores, see Online Appendix A. For fall ELA scores, the correlation between observed and projected RIT scores is 0.996, with an RMSE of 2.3 points.[13]

Figure 2 illustrates how even small changes in estimated scores using projection methods could have a large impact on estimating summer learning rates.[14] Figure 2 presents two hypothetical students as they progress through school between January 2008 and January 2012. Student 1's observed scores—and their test dates—are shown in orange. In dashed green, we project Student 1's achievement scores linearly based on their school-year learning rate. The green line connects the student's projected achievement on the last day of school to the projected achievement on the first day of school after that summer. In some grades, the summer learning gains estimated in the absence of school calendar information would be positive, but instead appear negative once the projections are used. The results are similar for Student 2 (red solid= observed scores, blue dashed= projected scores). The linear projection process—though it produces scores strongly correlated with observed scores—could have a profound impact on the estimated summer learning gain/loss. In this paper, we therefore use the projected RIT scores in favor of the observed RIT scores. However, in Online Appendix A, we reconduct the analyses using observed scores in place of projected scores and replicate the figures in this paper that capture the main findings.

Analytic Sample

For the current analysis, we first restrict the NWEA sample to students observed in grades 1 through 8 (because these are the grades with most complete coverage) and to the 89% of those students who neither repeat or skip grades. In our preferred models, we also restrict the sample to a "balanced panel"—that is, the subset of students who possess test scores for the full grade range being included in the model. For instance, if we examine test score patterns between 1st through 5th grade in a given model, only students who have both fall and spring test scores in every grade between 1st and 5th grade (that is, a full vector of all 10 reading test scores) will be included in the sample. While this is quite restrictive sample limitation, it ensures that our findings cannot be conflated with compositional changes from one time point to the next. In Online Appendix B, we replicate our primary findings on a less restrictive sample by running models with only 3 consecutive grades at a time (e.g., grades K through 2, grades 3 through 5, etc.). In these models, more students are included because the vector of required test scores is much shorter. These two samples have different advantages in terms of internal and external validity. Ultimately, however, results are relatively consistent (see Online Appendix B).

In Table 1, we compare the demographic descriptives for the students, schools, and districts from 4 groups: The population of U.S. public schools (from Common Core of Data), the entire population of NWEA test takers, the subset of students who meet the *less* restrictive inclusion criteria (for Online Appendix B), and the *more* restrictive inclusion criteria for our preferred results . See Table 1 (for simplicity, we conduct this comparison in the 2011-2012 school-year). First, recall that that a student-level indicator of free/reduced-price lunch (FRPL) status is not available in the NWEA dataset. However, at the school level, the mean percent of students in a school who are FRPL-eligible is very similar across the four groups: 50% both nationally and in NWEA universe of schools, 48% in the larger Online Appendix B sample, and 51% in the more restrictive, primary analytic sample. In many ways, the NWEA sample reflects the U.S. public school population. For instance, it is similar in terms of percentage of students identified as Black, Asian, White, and male. In addition, the majority of U.S. public schools are in rural geographic codes, followed by suburban and urban geographies, and this ordering also holds in NWEA. Many of the district characteristics are also quite similar.

To consider limitations to generalizability, we point out that the largest differences between the U.S. public school population and the NWEA universe are that (a) the NWEA sample has a lower percentage of Hispanic students, (b) the average NWEA school has somewhat smaller mean enrollment, and (c) the NWEA districts tend to have more schools in them, have a lower percentage of FRPL students, and are less likely to be rural. These differences could be connected to potential for *un*observable differences between the NWEA sample and the public-school population (e.g., orientation towards innovation and technology, resource allocation strategies, district leadership, etc.). What is also of note, however, is the sheer number of students in the NWEA universe in 2012 alone. NWEA students may comprise more than 11% of the entire K-12 public school population in 2012. NWEA data are available in nearly 37% of all U.S. public schools and in over half of all districts. This population is large enough to be of interest in its own right. Nonetheless, the lack of national representativeness is a weakness of NWEA data, relative to ECLS-K data.

Finally, we examine how the analytic sample limitations affect the characteristics of the NWEA students included in the models (compare the right three columns of Table 1[15]). The final column reflects the requirements for inclusion in the balanced panel. Generally, the analytic restrictions do not dramatically alter the descriptive profile of included NWEA students, schools or districts. However, the primary analytic sample has a higher percentage of white students than the NWEA full dataset (60% versus 53%), and the schools tend to be a smaller (mean enrollment of 391 versus 486) and are less likely to be suburban.

III. Methods

We use a multilevel model to estimate an individual learning trajectory for each student as they progress through sequential school-years and summers. We then look across students to estimate how much students tend to gain, on *average*, during the school-year versus what they typically lose during the summer. A multilevel modeling approach also allows us to estimate the *variation* in these gains/losses across students. Our multilevel model uses a Bayesian approach to estimate the variances and covariances. This approach produces more conservative estimates of student-level variances and is therefore preferable to calculating the raw standard deviation of summer gains, which reflects measurement error (Raudenbush & Bryk, 2002).[16]

Longitudinal Multi-Level Models

We use a two-level random effects (hierarchical) model, in which the outcome of interest is a test score, $Score_{ti}$, for student i at grade-semester t. In our preferred models, we separately model scores in 1st through 5th grade (students included here must have all 10 math score outcomes) and then in 5th through 8th grade[17] (again, students must have all 6 test scores in these grades). For brevity, we present the model (Eq. 1) for math scores from grade 6 through grade 8. These six repeated observations (L1) are nested within students (L2):

<u>Level One: Repeated Observations of Students (i) across Grade-Sems (t)</u>

$$Score_{ti} = \pi_{0i} + \pi_{1i}(schyr6_{ti}) + \pi_{2i}(sumaf6_{ti}) + \pi_{3i}(schyr7_{ti}) + \pi_{4i}(sumaf7_{ti}) + \pi_{5i}(schyr8_{ti}) + \pi_{6i}(sumaf8_{ti}) + \varepsilon_{ti} \qquad \text{where } \varepsilon_{ti} \sim N_{iid}(0, \sigma)$$

<u>Level Two: Students (i)</u>

$$\begin{aligned} \pi_{0i} &= \beta_{00} \\ \pi_{1i} &= \beta_{10} + r_{1i} && \text{where } r_{1i} \sim N_{iid}(0, \tau_{1,1}) \\ &\vdots \\ \pi_{6i} &= \beta_{60} + r_{6i} && \text{where } r_{6i} \sim N_{iid}(0, \tau_{6,6}) \end{aligned} \qquad \text{Eq (1)}$$

At L1, students' growth trajectories are modeled with a set of dummy variables—$schyr6_{ti}$, $sumaf6_{ti}$, $schyr7_{ti}$, $sumaf7_{ti}$, etc.—for each grade-semester. Each is coded 1 if the observation occurred *on or after the ending timepoint for the period*.[18] This coding scheme is different than that chosen in some prior work[19] and may at first seem confusing, but it has the advantage of giving the level-one coefficients intuitive meaning that now match the variable names: They represent an individual student *i*'s grade-specific school-year gain or grade-specific summer gain/loss. For example, π_{1i}—the coefficient on $schyr6_{ti}$—captures student i's 6th grade school-year learning gain. The coefficient on $sumaf6_{ti}$ captures student *i*'s summer after 6th grade gain/loss. These coefficients are now the very learning gains/losses we are interested in estimating for each student. We allow all of the level-one coefficients π_{1i} through π_{6i} to vary randomly at the student level, and we assume that the level-two errors (r_{1i} through r_{6i}) are normally distributed with a mean of zero and a constant variance given by $\tau_{1,1}$ through $\tau_{6,6}$. At level two, we use a fully unstructured covariance matrix, meaning that we estimate the variances of and correlations among all period-specific gain/losses rather than constraining them to be zero or any other known value. These models estimate the parameters we need to answer each of our research questions, in turn.[20]

IV. Results

(RQ1) Average Students' School-Year vs. Summer Learning Gains/Losses across Grades

We present findings both formally (i.e., point estimates in tables) and visually to make takeaways as tangible as possible. For instance, to address this first question, we present the β coefficients (or "fixed effects" in the language of HLM) in Table 2 (ELA) and Table 3 (math) because, substantively, they capture mean gains/losses in each grade and following summer. These β coefficients are also graphed in Figure 3 as mean growth trajectories.

During school-years. To contextualize the findings about summer experiences, we first discuss mean school-year learning gains. Beginning the left column of Table 2 (ELA), we find that students' school-year learning gains are largest in the early grades and generally diminish over time. This is depicted in Figure 3 with blue, dashed lines. For instance, students gain on average 23.7 ELA test score points in 1st grade, 18.5 points in 2nd grade, 13.3 points in 3rd grade, and so on. By 8th grade, the average ELA learning gain on NWEA's RIT scale is just 4.4 points. We observe a very similar pattern for math (left column of Table 3). In all grade levels, the average student gains—as opposed to loses—ground during school-years. This suggests that students accumulate knowledge over time during school-years as measured by the NWEA MAP test.

During summers. The pattern of mean summer learning gains/losses—the β coefficients in the right column of Table 2 and Table 3—are shown as solid red lines in Figure 3. Summer estimates differ from school-year gains in two important ways. First, in both ELA and math, the summer coefficients between 1st and 8th grade are negative and tend to be smaller in magnitude. For instance, the average ELA loss in the summer after 1st grade is -6.6 test score points, -3.9 in summer after 2nd,

-3.4 in the summer after 3rd, and falls to a low of -0.9 just before grade 8. In math, the mean summer learning estimates are also negative and of similar magnitude. An implication here is that, depending on grade, the average student loses between 17 and 28% of their school-year ELA gains (a 9-month period) during the following summer (a 3-month period). In math, the relative losses are a little larger: The average student loses between 25 and 34% of each school-year gain during the following summer.

The second way in which summer estimates differ from their school-year counterparts is that the magnitude of mean summer learning losses does not decrease over time to the same degree as school-year learning. Put differently, although mean school-year gains in ELA fall from 23.7 to 4.4 across grades, mean summer losses stay within a tighter range of -6.6 to -0.9.

Turning to the visual representation of these findings in Figure 3, we consistently see a zig-zag pattern at every grade level, though the intensity of gains/losses flattens at higher grades. These results generally confirm the notion that summers can be characterized as a time when, on average, students lose ground. Historically, SLL studies have not reached consensus on the direction of mean summer learning rates; some find losses, while others find stagnation, mere slowdown, or a mix of results across grades, subjects, or datasets. The current study joins those that find mean losses, but we will see that the 95% plausible value range across students always includes zero. However, we caution against over-emphasizing mean SLL since it will become clear that this mean does not well-characterize what students experience in the summer, because it masks dramatic underlying variability across students.

(RQ2) Variation in Students' School-Year vs. Summer Learning Gains/Losses, by Grade

It is important to recognize that the trends illustrated in Figure 3 only tell us one part of the story: the seasonal learning patterns for the *average* student. However, achievement disparities are driven by *differential* learning patterns, and so we now focus on how students vary on both school-year and summer learning gains/losses. We are particularly interested in determining whether student growth trajectories vary more during school-years or summers.

During school-years. We begin by examining variability in school-year learning across students. The first column of Table 2 (ELA) and Table 3 (math), contains the estimated standard deviations (SDs) of learning gains/losses across students in and after each grade (i.e., the square root of the diagonal elements of the tau matrix). For example, while we saw that the average student gains 23.7 ELA points in grade 1, students also typically differ from this mean by 9.7 points. To illustrate the magnitude of this variability, we construct a 95% plausible value range (PVR) for learning gains across students (under the assumption of normality (Raudenbush & Bryk, 2002)). These are reported in Table 2 (ELA) and Table 3 (math) beneath the corresponding student SD. To continue with the example of grade 1 ELA gains, we expect that 95% of students would have an average learning gain between 4.4 and 42.7 ELA test score points. Therefore, in 1st grade, students at the high end of the PVR gain about 80% more than the average student.

Estimates of the SD of school-year learning gains across students are relatively consistent across school-years and subjects, generally in the range of 6 to 10 test score points. In grades that exhibit smaller *average* school-year gains, this variation implies larger discrepancies across students. For instance, in 8th grade when average growth is only 4.4 test score points during the school-year, we see a 95% PVR across students of -7.0 to +15.9 points. Here, students at the top of this PVR will experience nearly four times larger gains than the average student. Students at the lowest end of that same PVR, however, are actually *losing* ground during 8th grade.

To juxtapose mean gains/losses with variation around them, we calculate the ratio of the variation (SD) across students for each learning gain to the mean learning gain. Larger ratios indicate greater variability, relative to the mean gain. In 1st grade ELA, that ratio is about 0.41 (9.7 over 23.7), indicating that the SD is a little less than half the size of the mean gain. In ELA, that ratio grows slowly across grades and reaches 1.3 in grade 8 (that is, the SD is now about 30% larger than the mean). The ratio also increases across grades in math, but less dramatically from 0.40 in 1st grade to 0.91 in 8th grade. However, the fact that the relative variability in learning gains grows as students progress through school may suggest that inequities in achievement accumulate to some extent during school-years as students who are underprepared are being left further and further behind with each successive grade.

During summers. While the variability in school-year patterns are interesting in and of themselves, our main interest lies in whether the summer gains/losses vary more than gains in the school-year periods. This has direct implications for our understanding of when discrepancies in student achievement arise across the course of students' school-age years. Turning to the second columns of Table 2 (ELA) and Table 3 (math), we see that the SD for a given summer tends to be a little smaller than the SD in the preceding school-year (with the exception of 1st grade). For instance, in 3rd grade math, the SD is 6.6 in the school-year and 3.6 in the following summer. This is expected; the summer is about one-third the length of the summer and so gains will be smaller. However, in a relative sense, the summer SDs are *much* larger with respect to means. In ELA, the SD-to-mean ratios described above are much larger in summers, ranging from 1.4 to as high as 5.2. A ratio of 5.2 indicates that the SD is over five times larger than the mean loss. Recall that the *largest* such ratio during a school-year was only 1.4. In math, we also see that summer ratios, which range from 0.8 to 2.3 are larger than school-year ratios (which only range from 0.40 to 0.91). Keep in mind that this larger summer variation is arising in a comparatively shorter time (around 9 versus 3 months). This highlights the fact that a great deal of variability in gains/losses is packed into a relatively short time frame.

The PVRs are large for summer learning loss. Take 2nd grade math as an example: Summer learning loss in grade 2 for math (second column of Table 3) ranges from -16.3 to +6.8. While students at the top of that PVR are gaining, during the summer, another 32% of average growth from the preceding 2nd grade school-year (6.8 over 18.6), students at the bottom of the PVR will *lose during the summer just as much as the typical student gained in 2nd grade*. Looking across all grades in ELA, we find that students at the top of the summer loss PVR will gain during the summer between 45 to 154% of the mean growth in the preceding grade (12 to 86% for math). However, students at the bottom of the summer loss PVR will *lose* during the summer between 93 to 194% of the mean growth in the preceding grade (73 to 136% for math). In sum, some students experience *accelerated* learning during the summer, relative to the preceding school-year, while others lose nearly all of their prior gains.

The takeaways for RQ2 are also illustrated visually in Figure 4 (ELA) and Figure 5 (math), wherein we present box plots of individual students' empirical Bayes estimated learning gains and losses in each school-year and summer. These concisely capture the essence of what is presented in the tables: Larger gains during school-years that diminish across grades, smaller average losses during summers that are more consistent in magnitude, but real variability around typical gains/losses. In Online Appendix B, we replicate Figure 4 (ELA) and Figure 5 (math) using results from models using a shorter three-grade increment. Though the data coverage is sparser before 1st and after 9th grade, we do include those grades in Online Appendix B.

In sum, students certainly appear to vary in terms of how much they learn during the school-year, but most students tend to exhibit some test score gains while in school. However, the picture in the summer is quite different. While our results re-document the *mean* summer learning loss phenomenon, this finding obscures a more problematic pattern: For mostly unknown reasons,[21] certain students can gain at a faster rate in the summer than the mean rate in the preceding school-year, while other students could lose most of what is typically gained.

(RQ3) Student-Level Correlation of Summer Gains/Losses across Summers

To this point, we have highlighted important variability in summer learning patterns across students. However, if that phenomenon occurs to students randomly—that is, a student might gain in one summer and then randomly lose in the next—then the contribution of summer learning loss to end-of-school achievement *disparities* would be limited. However, if the same students tend to experience losses summer after summer, while others gain summer after summer, it would lead to a more dramatic "fanning out" of student outcomes as they progress through school. We would be particularly concerned if the students who exhibit the greatest summer losses also tend to be from historically marginalized student populations—a question that has been taken up in many prior SLL studies. However, since student demographics appear to only account for about 4% of the variance in summer learning rates (von Hippel et al., 2018), we explore the systematicity of SLL across grades beyond just the differences by race and class.

To explore this question empirically, we examine from our multilevel models the estimated covariances of students' summer losses across grades.[22] The upper panels of Table 4 (ELA) and Table 5 (math) present these covariances (expressed in correlations). Positive correlations are the most problematic: summer losses accrue to the same students over time in a way that would contribute to the widening of end-of-schooling student outcomes. Correlations near zero would suggest gains/losses occur randomly. In ELA, all correlations are positive (between 0.12 and 0.65), and most are substantively large. The corresponding correlations are also positive in math, ranging between 0.10 and 0.65. This suggests that students who lose ground in the summer tend to also lose ground in other summers. Likewise, students who make summer gains in one summer are also more likely to make gains in other summers. While few other studies have presented similar correlations across summers, von Hippel et al. (2018) also find a positive (though weaker) relationship between summer learning rates in the summers after K and 1 for reading (of $+0.06$) in ECLS-K:2011 but interestingly find that relationship is negative (-0.21) in math.

In the lower panels of Table 4 (ELA) and Table 5 (math), we also present the correlations of summer gains with school-year gains. Given that we have observed a notable zig-zag pattern in learning trajectories and that the majority of students do exhibit learning *gains* while in school, we should anticipate that these correlations will be negative, particularly in adjoining periods (e.g., when a student loses ground in the summer after grade 4, they start grade 5 in the fall from a lower point from which to grow). Indeed, this is what we observe. For ELA, all but one[23] of the 16 correlations presented in the lower panel of Table 4 are negative, and correlations from adjoining periods are strongest. Of course, the more time separates the given summer (rows) and school year (columns), the weaker that negative relationship becomes. For instance, school year gains in grade 1 exhibit a negative correlation of -0.41 with summer gains/losses in the summer directly after grade 1, -0.23 with the summer after grade 2, -0.01 with the summer after grade 3, and $+0.01$ with the summer after grade 4. The results for math (lower panel of Table 5) follow a very similar pattern. These findings are also consistent with those of von Hippel et al. (2018) who also report negative correlations between summer and school year learning rates across grades K, 1, and 2 on the order of -0.55 to -0.21 in both reading and math.[24]

(RQ4) The Role of Summers in Producing End-of-school Outcome Disparities

Taken together, these three findings—(RQ1) slightly negative mean summer losses, (RQ2) large variances in summer loss/gains, and (RQ3) systematic gain/loss patterns across summers—imply that end-of-school achievement disparities arise at least partly during the summer. How large of a role do summers play? To consider this question, we begin by presenting a thought experiment designed to characterize the role of summers between grade 1 and 8. We imagine a hypothetical scenario in which all students enter 1st grade at the exact same achievement level, and all students experience the exact same (let’s say, the mean) learning gain in each grade while school is in session. If there were no summer periods, all students in this scenario would end 8th grade with the same test score, because no variation in gains arises while in school. We now return to the results from our multilevel model to characterize three plausible student experiences during the summers following each grade: The typical gain among students in the top, middle, and bottom thirds of a given summer’s gain/loss distribution.[25] We now illustrate these three levels of summer experiences in Figure 6 (ELA in top panel, math in bottom panel), while assuming school-year gains are always equal (i.e., parallel slopes of dashed blue lines fall to spring).

Figure 6 shows how the differences in summer experiences *by themselves* would lead to sizeable achievement over time. In ELA, the spread in test scores at the end of 8th grade is from about 185 to 255 test score points (and about 200 to 265 in math)—around 2.5 standard deviations of spring 8th grade RIT scores. This thought experiment illustrates the idea that, even in an ideal world where school inequities could be eliminated, achievement disparities would arise simply because of the summer break. The “fanning out” of achievement during these school-age years would need to be addressed in large part with respect to summer experiences.

V. Conclusion

Reflections on Findings

In this paper, we conduct a thorough exploration of the seasonality of learning from a dataset covering nearly 18 million students in 2008 through 2016 across all 50 states. We focus on characterizing the degree of variability in students’ summer experiences and the role of summers in contributing to end-of-school achievement disparities. We find that students, on average, do indeed lose meaningful ground during the summer period in both math and ELA.

We add to the existing research by estimating the total variance across students in SLL. For instance, consider the SLL pattern after second grade, in which the average school-year gain is 18.6 points in math. During the summer that follows, the 95% plausible value range indicates that some students will *lose* as much as 16.3 test score points in math during summer, while other students could *gain* up to 6.8 test score points (relative to a mean SLL of 4.8 points). Students do also exhibit significant variance in school-year learning, however the lower bounds of the 95% plausible value ranges during the school-year tend to be much closer to zero. This means that, while some students learn more than others during the school-year, most students are moving in the same direction—that is, making learning *gains*—while school is in session.

The same cannot be said for summers. During the summer, a little more than half of students exhibit summer learning losses, while the other half exhibits summer learning gains. It is clear that the summer period is a particularly variable time for students. We find that some students can in fact maintain average school-year learning rates during the summer in the absence of formal schooling. Other students, however, will lose nearly as much as what is typically gained in the preceding school-year.

This remarkable variability in summer learning appears to be an important contributor to widening achievement disparities during the school-age years. However, most education research tends to overlook the summer period by focusing on programs, policies, and practices designed to shape *schooling* experiences. But summers deserve greater attention: In Figure 7, we present the distribution, across students, in the percentage of their absolute value fluctuations from 1st through 5th grade that occurs during summers. One can think of this as the percentage of each student's up/down "pathway" between their initial and end score that arises during the summer. Far from having no role in outcome inequality, we see that on average, 19.4% of students' ELA test score changes occur during the summer (19.3 for math).[26] However, for some students, summer fluctuations account for much more—even upwards of 30%—of where they end up.

Our findings also suggest that summer learning gains/losses can be quite large and may accrue non-randomly across students. If the likelihood of experiencing a loss during the summer were independent across students and grades, we would expect that only 24% of students would exhibit losses in five consecutive summers.[27] In contrast, we actually find that 52% exhibit losses (in ELA) in all 5 consecutive years observed—more than double what one would expect by chance. Furthermore, the average student in this group ultimately loses 39% of their total school-year ELA gains during the summer periods (results are similar for math). This suggests that negative summer decrements tend to accumulate to the same students over time and that these consecutive losses add up a sizeable impact on where students end up in the achievement distribution.

Contextualizing Findings in Larger Body of SLL Literature

Historically, SLL studies have not reached consensus on the direction (+/-) of mean SLL. Some find mean summer learning *losses* (e.g., Allinder et al., 1992; Borman et al., 2005), summer learning *stagnation* (e.g., Benson & Borman, 2010; Downey et al., 2008), summer learning *slowdown* (e.g., Alexander et al., 2001; Burkam et al., 2004; Quinn et al., 2016), or a mix of the three (e.g., Downey et al., 2004; Heyns, 1978; von Hippel et al., 2018). For instance, von Hippel et al. (2018) finds positive summer learning rates in some grades, subjects, or ECLS-K cohorts, but flat or negative rates in others. The current study joins those that find mean summer *losses*. We observe this in every summer between 1st and 8th grade in both math and ELA.

How does the consistency we see across subjects align with not only recent studies, but also with Cooper et al. (1996)'s meta-analysis which found, on average, more negative impacts of summer vacation in math-related subjects than in reading-related subjects? Cooper et al. (1996) hypothesize that math skills are more the domain of formal schooling, while reading happens both at home and in school. However, the authors also point out that SLL skill patterns did not always fall along a math/ELA divide: Rather, the skills they viewed as more "procedural" (e.g., spelling and math computation) declined most during the summer (although reading comprehension also appears to decline in summers, which does not align with this theory). Since we cannot disaggregate our results to more specific math and reading skills, it is less clear whether our findings are in conflict with those of Cooper et al. (1996). Moreover, while Cooper et al. found patterns of skill-specific gains/losses, in more recent studies that document mean SLL, no clear pattern by subject has emerged.[28]

Is the magnitude of mean SLL similar across studies? As a reminder, the current study covers different grade levels than the ECLS-K studies; the only overlapping summer is the one after 1st grade (see Figure 1 to review which studies cover which summer grades). This may be partly responsible for any disparate findings. However, in this case (summer after 1st grade), we think the results from NWEA and ECLS-K:11 are complementary. Take seasonality in ELA learning as an example: von Hippel et al. (2018) document a modest but statistically significant mean SLL rate of -0.02 SDs/month. We find a mean SLL rate of around -2.2 points/month, with a 95% plausible value range across students that includes zero (for context, the K fall SD is about 13 points). However, once these mean SLL rates are contextualized with respect to the student level SDs in SLL, the studies look even more similar: both show that the student SD is much larger—2 to 4 times larger— than the mean of SLL.

Most prior SLL research has focused on SES or racial/ethnic gaps in SLL, which is not the focus of the current study. As highlighted in Figure 1, we are aware of only one other study that examines seasonal patterns of unconditional variance in SLL.[29] Our results support two primary claims: (1) First, we find that variation in achievement grows significantly from grade 1 to 8. (2) Second, summer learning varies dramatically and relatively more so than school-year learning.

With respect to the first claim, while we find evidence of widening achievement disparities when we follow students from grade 1 to 8, prior research has not reached consensus on this matter. Claessens, Duncan, and Engel (2009) used the IRT-based scale score versions of achievement from ECLS-K:99 and document SDs that grow from K to 8 by 141%.[30] Test score scaling appears to be crucial in this debate, however, because when von Hippel, Workman, and Downey (2018) use improved, IRT-based theta achievement measures to report grade-specific SDs of scores, they actually find that those standard deviations *shrink* from K to 2. Despite the fact that both the current study and von Hippel et al. (2018) use vertically-scaled scores, the former indicates variation grows, while the latter suggests variation may shrink.

This debate about whether or not achievement disparities widen as students move through school is long-standing. It may seem counterintuitive that, as students move through school experiencing both different schools and different summers, their achievement would become more homogenous. But again, test score scaling will prove central to this question. Vertically-scaled scores are probably the appropriate *theoretical* approach to measuring growth over time, yet because the assumptions of vertical scales are hard to verify, it is difficult to conclude that a given scoring technique indeed yields the "right" scores. Vertically-scaled scores, too, can suffer from measurement artifacts (e.g., scale shrinkage or ceiling effects). Camilli, Yamamoto, and Wang (1993) capture the conundrum succinctly : "It cannot be determined whether developmental scales should show expansion or contraction. The criteria for determining useful vertical scales constitute a controversial topic of debate and research" (387).

Though not directly related to widening *unconditional* variance across grades, it is also useful to consider whether other researchers have found that race/ethnicity or SES gaps widen as students move through school, since demographic gaps could at least partly contribute to overall variation. Again, prior evidence is mixed. For instance, Duncan and Magnuson (2011) show increasing SES, Black-White, Hispanic-White, and gender achievement gaps in math between 1st and 5th grade. In Reardon (2008), IRT-based theta scores show the Black-White gap increases from -0.32 in grade K to -0.41 in grade 5. Recent results based on ECLS-K:11 from von Hippel and Hamrock (2019) and Quinn et al. (2016) both suggest that the Black-White gap may grow in the early grades, but—in contrast to prior studies that may suffer from measurement artifacts—SES gaps may shrink between K and 2.

With regard to our second claim—that summers contributes more to achievement disparities than school-years—our results are consistent with the one other study in this domain (von Hippel et al., 2018). In both studies, students exhibit meaningful student-level variation in both school-year and summer learning. But, as in the current study, von Hippel et al. (2018) find that the student-level SDs of learning rates are larger in the summers. They find this in both ECLS-K cohorts, both subjects, and in the summer after K and 1. Though school-years are generally three times longer than summers and thus have more opportunity to contribute to widening achievement disparities, summers clearly play a key role in where students end up.

Finally, we can provide some limited reflections on the recent debate about whether inferences concerning the growth and seasonality of SES or race gaps have been distorted by measurement artifacts in earlier work. The von Hippel and Hamrock (2019) article highlights the importance of scaling: The same dataset can yield opposing inferences when a different version of the test scores is used. While we find the arguments made by von Hippel and Hamrock (2019) regarding preferred measurement properties compelling, we do not have the ability in the current dataset to empirically explore these issues, since we do not have item-level data. Moreover, their study documents a different phenomenon—race and SES gaps—than we document here. We should not necessarily expect that the patterns in overall variability in SLL would move in tandem with patterns by demographics, since demographics seem to explain only a little of the variation in SLL. Whether or not this is an appropriate interpretation of von Hippel and Hamrock's findings, their study has shaken some people's confidence in the idea that SLL matters. However, as in von Hippel and Hamrock (2019), we use too use vertically-scaled test scores and still find clear evidence that SLL exists and contributes substantially to where children end up in the achievement distribution. This suggests that SLL is very much worthy of continued research.

Study Limitations

First and foremost, the NWEA dataset does not include key variables to explore SLL gaps (e.g., FRPL, language, special education status, links to teachers). In addition, the current study rests on the assumption that NWEA's RIT scores are a valid measure of student math and reading in both the fall and spring and over time (i.e., the vertical scaling). NWEA's MAP test is a formative assessment without stakes, and it is not entirely clear that there are incentives in place for students and teachers to take it equally seriously in the fall and spring. Students tend to spend slightly less time on the fall tests than their spring tests. One would be concerned if this signals that students do not put forward as much effort on their fall assessments, thus making summer learning losses appear larger than they actually are. We believe that the difference in time spent is not large (about 6 additional seconds per item, on average, in the spring), and we find that controlling for time spent on test affects the results very little. In addition, most of the analyses herein do not rely on making direct comparisons across distal grades, thus reducing reliance on vertical scaling properties for these particular inferences. That said, the findings herein should be considered with these caveats in mind.

Implications

Our results show that summers contribute more to achievement disparities than school-years. Our findings to this effect align with prior work (e.g., Downey et al., 2004; von Hippel et al., 2018), though the current study provides perhaps the most comprehensive empirical analyses to date, given its large sample, extension beyond the early grades, and its focus on overall variation.

This finding has implications for outcome inequality, yet it can be viewed through two different lenses. On the one hand, it can be interpreted for what it says about *summers*. These periods, it seems, are more relevant for the expansion of outcome variation. Some will find themselves looking to summers as a time for intervention, and perhaps even questioning whether long summer breaks should be standard practice.

On the other hand, this finding can be interpreted for what it says about the *school-year*—that is, how we understand the role of schools in the production of outcome inequality. The summer can be thought of as a counterfactual to schooling, giving us a window into how inequality would grow in the absence of schools' influence. SLL researchers have pointed out that, if learning rates vary less during the school year than during the summer, schools may be countering some of the powerful forces that exacerbate inequality when school is not in session.

Should schools be reframed, then, as 'equalizers'—ameliorating rather than exacerbating outcome inequality? Certainly, this perspective is not widely embraced in the education research community. It *is* still true that, during school-years, some students gain much more than others. Perhaps, then, it would be more precise to say that schools may not intensify inequality but also cannot fully counter it, nor even hold it constant. In a sense, this question is a philosophical one that depends on what one thinks the purpose of public schooling is.

We motivate the current study based on a lack of consensus across prior SLL research, along with recent questions about measurement artifacts in foundational studies. Our goal is to conduct basic research to clarify our understanding of this important phenomenon. Since we focus more on surfacing just how varied summer learning is and how little we understand about it, making specific policy recommendations is premature. Below, we offer our thoughts about potential directions for future applied research.

Since our results show that achievement disparities widen during school-years, we should continue to develop policies that change how students experience schools, particularly on issues of access. Yet, even in a hypothetical scenario where students all learn the same amount during the school-year, the time spent out of school in summer break, by itself, gives rise to much of the dramatic spread of achievement outcomes, on the order of several standard deviations.

One natural question, then, is whether to extend the school-year to reduce summer atrophy and minimize opportunities for this divergence to occur. However existing research on year-round school calendars does not indicate that SLL is mitigated by these schedules (Graves, 2011; McMullen & Rouse, 2012). It is possible that year-round calendars implemented to address over-crowding (a common impetus) may have different impacts on learning than year-round calendars implemented explicitly to reduce SLL, but to our knowledge this hypothesis has not been tested.

Another policy lever might be to focus on programs that bridge the gap between May and August like summer school. The causal evaluation of summer school is often fraught, given the non-random selection of who is required to enroll and known issues around low attendance (especially in higher grades). Yet there is growing evidence that summer interventions can help mitigate students' SLL (Kim & Quinn, 2013; McCombs et al., 2012; McCombs et al., 2015). For instance, seven New Mexico school districts randomized early grade children in low-income schools into an ambitious (and presumably expensive) summer program called K-3+, that essentially amounted to a full-blown extension of the typical school-year for much of the summer period. Early results from the experimental study indicated that children assigned to K-3+ exhibited stronger literacy outcomes across four domains of the Woodcock Johnson achievement assessment (Cann, Karakaplan, Lubke, & Rowland, 2015).

Our results *also* suggest that we should look beyond schooling solutions to address out-of-school learning disparities. Researchers have pointed to differential resources in terms of families' economic capital, parental time availability, and parenting skill and expectations as potential drivers of outcome inequality (see for example, Borman et al. (2005)). Many of these resource differences are likely exacerbated by summer break when, for some families, work schedules come in greater conflict with reduced childcare. Many social policies other than public education touch on these crucial resource inequalities and thus could help reduce summer learning disparities.

Next Steps for SLL

We document the magnitude of a social problem—the role of summers in the growth of achievement inequality. While we can conclude that this happens and to what extent, the current dataset is not well-positioned to understand *why* summer learning patterns are so varied across students. Though it is an important first step to know when inequality arises and how unequal the learning patterns are, the obvious next question is: What accounts for that variation?

In some sense we have reached a precipice on SLL research. It seems clear that summers play a key role in outcome inequality and that the range of students' summer learning experiences is sizeable. Prior research suggests that this variability may fall partly along racial and socioeconomic lines (Alexander et al., 2001; Benson & Borman, 2010; Borman et al., 2005; Burkam et al., 2004; Downey et al., 2004; Gershenson, 2013; Heyns, 1978; Quinn, 2014; Quinn et al., 2016; von Hippel et al., 2018). However, prior research has also shown that demographic factors only account for a small part of the story here: In an insightful SLL study by Burkam et al. (2004) using ECLS-K:1999 data, the authors leverage the parent surveys of children's home and summer activities, in conjunction with student gender, racial, and socio-economic demographics—that is, most of the first-order candidates for explaining variability. However, they can explain only about 13% of the variance in learning gains in the summer after K. New research is needed to reconcile the fact that summer learning differs dramatically from child to child, but to date we have only limited insight into what accounts for most of that variation.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis, 23*(2), 171.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review, 72*(2), 167.

Allinder, R. M., Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1992). Effects of summer break on math and spelling performance as a function of grade level. *The Elementary School Journal*, 451-460.

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.

Borman, G. D., Benson, J., & Overman, L. T. (2005). Families, schools, and summer learning. *The Elementary School Journal, 106*(2), 131-150.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50*(2), 204-226.

Briggs, D. C., & Dadey, N. (2015). Making sense of common test items that do not get easier over time: Implications for vertical scale designs. *Educational Assessment, 20*(1), 1-22.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*(6), 551-576.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

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.

Camilli, G., Yamamoto, K., & Wang, M.-m. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388.

Cann, D., Karakaplan, M., Lubke, M., & Rowland, C. (2015). *New Mexico StartSmart K-3 Plus Validation Study: Evaluator's Report* Retrieved from Presented at APPAM Conference 2014.: http://ccpi.unm.edu/sites/default/files/publications/EvaluatorReport.pdf

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 28*(4), 415-427.

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.

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.

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. *Whither opportunity*, 47-70.

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.

Entwisle, D. R., & Alexander, K. L. (1994). Winter setback: The racial composition of schools and learning to read. *American Sociological Review*, 446-460.

Fitzpatrick, M. D., Grissmer, D., & Hastedt, S. (2011). What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review, 30*(2), 269-279.

Gershenson, S. (2013). Do summer time-use gaps vary by socioeconomic status? . *American Educational Research Journal, Forthcoming*, .

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

Gilkerson, J., & Richards, J. A. (2009). The power of talk. *Impact of adult talk, conversational turns and TV during the critical 0-4 years of child development: Boulder, CO: LENA Foundation*.

Graves, J. (2011). Effects of year-round schooling on disadvantaged students and the distribution of standardized test performance. *Economics of Education Review, 30*(6), 1281-1305.

Heyns, B. (1978). *Summer learning and the effects of schooling*: Academic Press New York.

Kaushal, N., Magnuson, K., & Waldfogel, J. (2011). *How is family income related to investments in children's learning?* : Russell Sage Foundation.

Kim, J. S., & Quinn, D. M. (2013). The Effects of Summer Reading on Low-Income Children's Literacy Achievement From Kindergarten to Grade 8 A Meta-Analysis of Classroom and Home Interventions. *Review of Educational Research*.

Kornrich, S., & Furstenberg, F. (2013). Investing in Children: Changes in Parental Spending on Children, 1972–2007. *Demography, 50*(1), 1-23.

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*: ERIC.

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Educational Research Journal, 41*(1), 115-157.

McCombs, J. S., Augustine, C., Schwartz, H., Bodilly, S., McInnis, B., Lichter, D., & Cross, A. B. (2012). Making Summer Count: How Summer Programs Can Boost Children's Learning. *Education Digest: Essential Readings Condensed for Quick Review, 77*(6), 47-52.

McCombs, J. S., Pane, J. F., Augustine, C. H., Schwartz, H. L., Martorell, P., & Zakaras, L. (2015). First Outcomes from the National Summer Learning Study.

McMullen, S. C., & Rouse, K. E. (2012). The Impact of Year-Round Schooling on Academic Achievement: Evidence from Mandatory School Calendar Conversions. *American Economic Journal: Economic Policy, 4*(4), 230-252.

Quinn, D. M. (2014). Black-White Summer Learning Gaps: Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis*. doi:10.3102/0162373714534522

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

Quinn, D. M., & Le, Q. T. (2018). Are We Trending to More or Less Between-Group Achievement Inequality Over the School Year and Summer? Comparing Across ECLS-K Cohorts. *AERA Open, 4*(4), 1-19. doi:10.1177/2332858418819995

Rambo-Hernandez, K. E., & McCoach, D. B. (2015). High-achieving and average students' reading growth: Contrasting school and summer trajectories. *The Journal of Educational Research*, *108*(2), 112-129.

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

Reardon, S. F. (2008). Thirteen ways of looking at the black-white test score gap.

Skibbe, L. E., Grimm, K. J., Bowles, R. P., & Morrison, F. J. (2012). Literacy growth in the academic year versus summer from preschool through second grade: Differential effects of schooling across four skills. *Scientific Studies of Reading, 16*(2), 141-165.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. Retrieved from https://www.nwea.org/content/uploads/2018/01/2015-MAP-Norms-for-Student-and-School-Achievement-Status-and-Growth.pdf

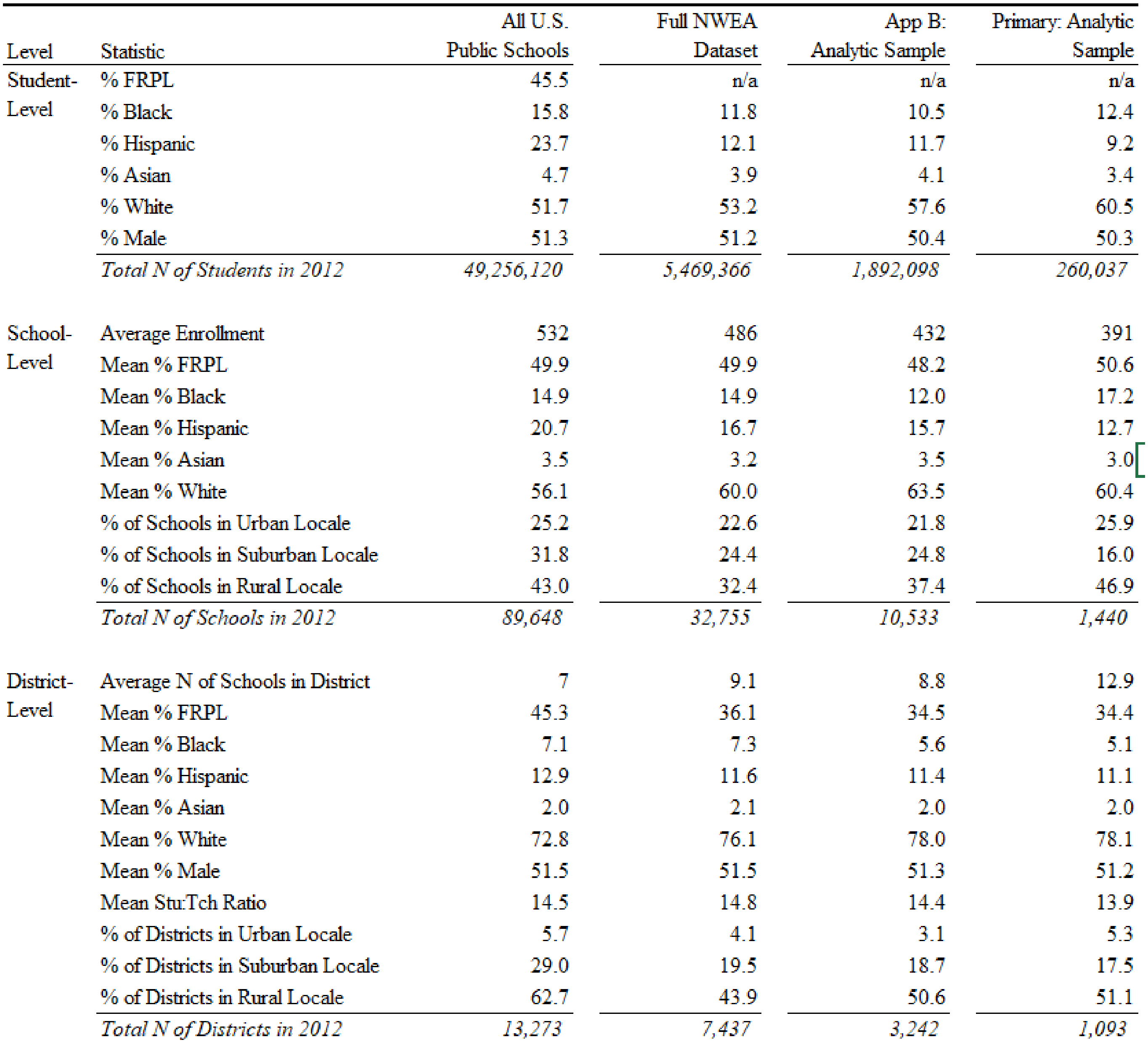
Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). *Combined User's Manual for the ECLS-K Eighth-Grade and K–8 Full Sample Data Files and Electronic Codebooks* Retrieved from Washington, DC: https://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf

von Hippel, P. T., & Hamrock, C. (2019). Do test score gaps grow before, during, or between the school years? Measurement artifacts and what we can know in spite of them. *Sociological Science*, *6*, 43-80.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are schools the great equalizer?". *Sociology of Education, 91*(4), 323-357.

Tables

Table 1. Descriptive Statistics in the Nation, in Full Dataset, in Analytic Sample in 2011-12



Level	Statistic	All U.S. Public Schools	Full NWEA Dataset	App B: Analytic Sample	Primary: Analytic Sample
Student-Level	% FRPL	45.5	n/a	n/a	n/a
	% Black	15.8	11.8	10.5	12.4
	% Hispanic	23.7	12.1	11.7	9.2
	% Asian	4.7	3.9	4.1	3.4
	% White	51.7	53.2	57.6	60.5
	% Male	51.3	51.2	50.4	50.3
	Total N of Students in 2012	*49,256,120*	*5,469,366*	*1,892,098*	*260,037*
School-Level	Average Enrollment	532	486	432	391
	Mean % FRPL	49.9	49.9	48.2	50.6
	Mean % Black	14.9	14.9	12.0	17.2
	Mean % Hispanic	20.7	16.7	15.7	12.7
	Mean % Asian	3.5	3.2	3.5	3.0
	Mean % White	56.1	60.0	63.5	60.4
	% of Schools in Urban Locale	25.2	22.6	21.8	25.9
	% of Schools in Suburban Locale	31.8	24.4	24.8	16.0
	% of Schools in Rural Locale	43.0	32.4	37.4	46.9
	Total N of Schools in 2012	*89,648*	*32,755*	*10,533*	*1,440*
District-Level	Average N of Schools in District	7	9.1	8.8	12.9
	Mean % FRPL	45.3	36.1	34.5	34.4
	Mean % Black	7.1	7.3	5.6	5.1
	Mean % Hispanic	12.9	11.6	11.4	11.1
	Mean % Asian	2.0	2.1	2.0	2.0
	Mean % White	72.8	76.1	78.0	78.1
	Mean % Male	51.5	51.5	51.3	51.2
	Mean Stu:Tch Ratio	14.5	14.8	14.4	13.9
	% of Districts in Urban Locale	5.7	4.1	3.1	5.3
	% of Districts in Suburban Locale	29.0	19.5	18.7	17.5
	% of Districts in Rural Locale	62.7	43.9	50.6	51.1
	Total N of Districts in 2012	*13,273*	*7,437*	*3,242*	*1,093*

FN: Data for U.S. public school population come from the NCES Common Core of Data and has been restricted to public schools (https://nces.ed.gov/ccd/). FRPL status is not available at the student level in the NWEA dataset. The Online Appendix B sample includes more NWEA students because it does not require students to have as long of a panel of available test scores to be included. The primary analytic sample used in the main narrative requires students to have up to ten available test scores in a row without missing data.

Table 2. ELA: HLM Model-Based Estimates of School-Year & Summer Learning Gains/Losses, Student-Level Standard Deviations, 95% Plausible Value Ranges across Students

		Model-Based Estimates		Post-Hoc Statistics for Given Grade					
		Gains/ Losses during the School Year	Gains/ Losses during the Following Summer	Means: % of Schyr Gain Lost in Summer	% More Gained @ Top of PVR in Schyr	Schyr: Ratio of SD to Mean Gain	Summer: Ratio of SD to Mean Gain	Summer: % of SY Gained @ Top of PVR	Summer: % of SY Lost @ Low of PVR
Grade 1	*coeff (beta)*	23.7 ***	-6.6 ***	28%	80%	0.41	1.6	55%	114%
	(se of beta)	(0.05)	(0.05)						
	stud sd (tau)	9.7 ***	10.4 ***						
	(stud 95% PVR)	(4.6 to 42.7)	(-26.9 to 13.7)						
Grade 2	*coeff (beta)*	18.5 ***	-3.9 ***	21%	109%	0.56	1.7	49%	93%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	10.3 ***	6.8 ***						
	(stud 95% PVR)	(-1.6 to 38.7)	(-17.2 to 9.3)						
Grade 3	*coeff (beta)*	13.3 ***	-3.4 ***	26%	119%	0.61	1.4	45%	98%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	8.1 ***	4.9 ***						
	(stud 95% PVR)	(-2.6 to 29.3)	(-13.0 to 6.3)						
Grade 4	*coeff (beta)*	10.1 ***	-2.6 ***	26%	132%	0.67	1.8	59%	118%
	(se of beta)	(0.04)	(0.04)						
	stud sd (tau)	6.8 ***	4.7 ***						
	(stud 95% PVR)	(-3.2 to 23.4)	(-11.9 to 6.7)						
Grade 5	*coeff (beta)*	7.8 ***	-2.2 ***	28%	204%	1.04	2.5	103%	169%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	8.1 ***	5.6 ***						
	(stud 95% PVR)	(-8.1 to 23.8)	(-13.2 to 8.8)						
Grade 6	*coeff (beta)*	6.4 ***	-1.6 ***	25%	236%	1.20	3.3	125%	186%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	7.7 ***	5.3 ***						
	(stud 95% PVR)	(-8.7 to 21.4)	(-11.9 to 8.8)						
Grade 7	*coeff (beta)*	5.2 ***	-0.9 ***	17%	275%	1.40	5.2	154%	194%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	7.3 ***	4.7 ***						
	(stud 95% PVR)	(-9.1 to 19.6)	(-10.1 to 8.4)						
Grade 8	*coeff (beta)*	4.4 ***		n/a	258%	1.32	n/a	n/a	n/a
	(se of beta)	(0.04)							
	stud sd (tau)	5.8 ***							
	(stud 95% PVR)	(-7.0 to 15.9)							

FN: We report Huber-corrected standard errors for the estimated beta coefficients, however due to the large sample sizes, all of the beta coefficients are highly statistically significant (distinguishable from zero). We focus more on the substantive significance more than the statistical significance in our discussion of these results.

Table 3. Math: HLM Model-Based Estimates of School-Year & Summer Learning Gains/Losses, Student-Level Standard Deviations, 95% Plausible Value Ranges across Students

		Model-Based Estimates		Post-Hoc Statistics for Given Grade					
		Gains/ Losses during the School Year	Gains/ Losses during the Following Summer	Means: % of Schyr Gain Lost in Summer	% More Gained @ Top of PVR in Schyr	Schyr: Ratio of SD to Mean Gain	Summer: Ratio of SD to Mean Gain	Summer: % of SY Gained @ Top of PVR	Summer: % of SY Lost @ Low of PVR
Grade 1	*coeff (beta)*	24.0 ***	-6.4 ***	27%	91%	0.46	1.7	58%	114%
	(se of beta)	(0.05)	(0.05)						
	stud sd (tau)	11.1 ***	10.7 ***						
	(stud 95% PVR)	(2.2 to 45.9)	(-27.4 to 14.6)						
Grade 2	*coeff (beta)*	18.6 ***	-4.8 ***	26%	92%	0.47	1.2	32%	88%
	(se of beta)	(0.04)	(0.04)						
	stud sd (tau)	8.7 ***	5.9 ***						
	(stud 95% PVR)	(1.6 to 35.6)	(-16.3 to 6.8)						
Grade 3	*coeff (beta)*	16.5 ***	-4.6 ***	28%	78%	0.40	0.8	12%	73%
	(se of beta)	(0.04)	(0.03)						
	stud sd (tau)	6.6 ***	3.7 ***						
	(stud 95% PVR)	(3.6 to 29.4)	(-12.0 to 2.7)						
Grade 4	*coeff (beta)*	14.2 ***	-4.3 ***	30%	86%	0.44	1.1	28%	96%
	(se of beta)	(0.04)	(0.03)						
	stud sd (tau)	6.2 ***	4.7 ***						
	(stud 95% PVR)	(2.0 to 26.3)	(-13.6 to 4.9)						
Grade 5	*coeff (beta)*	11.7 ***	-4.0 ***	34%	136%	0.69	1.3	51%	121%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	8.1 ***	5.2 ***						
	(stud 95% PVR)	(-4.2 to 27.5)	(-14.2 to 6.2)						
Grade 6	*coeff (beta)*	9.8 ***	-2.7 ***	28%	144%	0.73	1.8	61%	127%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	7.2 ***	4.9 ***						
	(stud 95% PVR)	(-4.4 to 23.9)	(-12.4 to 6.9)						
Grade 7	*coeff (beta)*	8.1 ***	-2.0 ***	25%	179%	0.91	2.3	86%	136%
	(se of beta)	(0.05)	(0.04)						
	stud sd (tau)	7.4 ***	4.6 ***						
	(stud 95% PVR)	(-6.4 to 22.6)	(-11.0 to 7.0)						
Grade 8	*coeff (beta)*	6.5 ***		n/a	163%	0.83	n/a	n/a	n/a
	(se of beta)	(0.04)							
	stud sd (tau)	5.4 ***							
	(stud 95% PVR)	(-4.1 to 17.2)							

FN: We report Huber-corrected standard errors for the estimated beta coefficients, however due to the large sample sizes, all of the beta coefficients are highly statistically significant (distinguishable from zero). We focus more on the substantive significance more than the statistical significance in our discussion of these results.

Table 4. ELA: Student-Level Correlations of Estimated Summer Gains with both School Year Gains and Summer Gains in other Grades

Corr(Summer, Summer Gains) across grades

Summer After -->	Grade 1, Sum.	Grade 2, Sum.	Grade 3, Sum.	Grade 4, Sum.	Grade 5, Sum.	Grade 6, Sum.	Grade 7, Sum.
Sum. After Grade 1	*1.00*						
Sum. After Grade 2	0.65	*1.00*					
Sum. After Grade 3	0.28	0.57	*1.00*				
Sum. After Grade 4	0.20	0.25	0.56	*1.00*			
Sum. After Grade 5					*1.00*		
Sum. After Grade 6					0.54	*1.00*	
Sum. After Grade 7					0.12	0.57	*1.00*

Corr(Summer & School Year Gains) across grades

School Year -->	Grade 1, SY	Grade 2, SY	Grade 3, SY	Grade 4, SY	Grade 5, SY	Grade 6, SY	Grade 7, SY
Sum. After Grade 1	-0.41						
Sum. After Grade 2	-0.23	-0.51					
Sum. After Grade 3	-0.01	-0.17	-0.63				
Sum. After Grade 4	0.01	-0.12	-0.19	-0.53			
Sum. After Grade 5					-0.61		
Sum. After Grade 6					-0.09	-0.58	
Sum. After Grade 7					-0.07	-0.08	-0.66

FN: In this table, we present the relevant off-diagonal elements of the covariance matrix, in the units of correlations. The model is run separately on early grades and later grades. Because the panel is only 9 years long, very few (less than 1%) of students have all 19 test scores between first through eighth grades. We therefore cannot estimate correlations across these two models.

Table 5 Math: Correlation Matrix Across Students' Summers Losses

Corr(Summer, Summer Gains) across grades

Summer After -->	Grade 1, Sum.	Grade 2, Sum.	Grade 3, Sum.	Grade 4, Sum.	Grade 5, Sum.	Grade 6, Sum.	Grade 7, Sum.
Sum. After Grade 1	*1.00*						
Sum. After Grade 2	0.65	*1.00*					
Sum. After Grade 3	0.15	0.43	*1.00*				
Sum. After Grade 4	0.09	0.15	0.49	*1.00*			
Sum. After Grade 5					*1.00*		
Sum. After Grade 6					0.42	*1.00*	
Sum. After Grade 7					0.10	0.53	*1.00*

Corr(Summer & School Year Gains) across grades

School Year -->	Grade 1, SY	Grade 2, SY	Grade 3, SY	Grade 4, SY	Grade 5, SY	Grade 6, SY	Grade 7, SY
Sum. After Grade 1	-0.56						
Sum. After Grade 2	-0.38	-0.57					
Sum. After Grade 3	-0.08	-0.14	-0.60				
Sum. After Grade 4	-0.06	-0.13	-0.10	-0.40			
Sum. After Grade 5					-0.68		
Sum. After Grade 6					-0.08	-0.59	
Sum. After Grade 7					-0.07	-0.09	-0.72

FN: In this table, we present the relevant off-diagonal elements of the covariance matrix, in the units of correlations. The model is run separately on early grades and later grades. Because the panel is only 9 years long, very few (less than 1%) of students have all 19 test scores between first through eighth grades. We therefore cannot estimate correlations across these two models.

Figures

Figure 1. Compare Studies: Datasets, Data Years, Grades Included, Number of Students, Location (page 1 of 2)

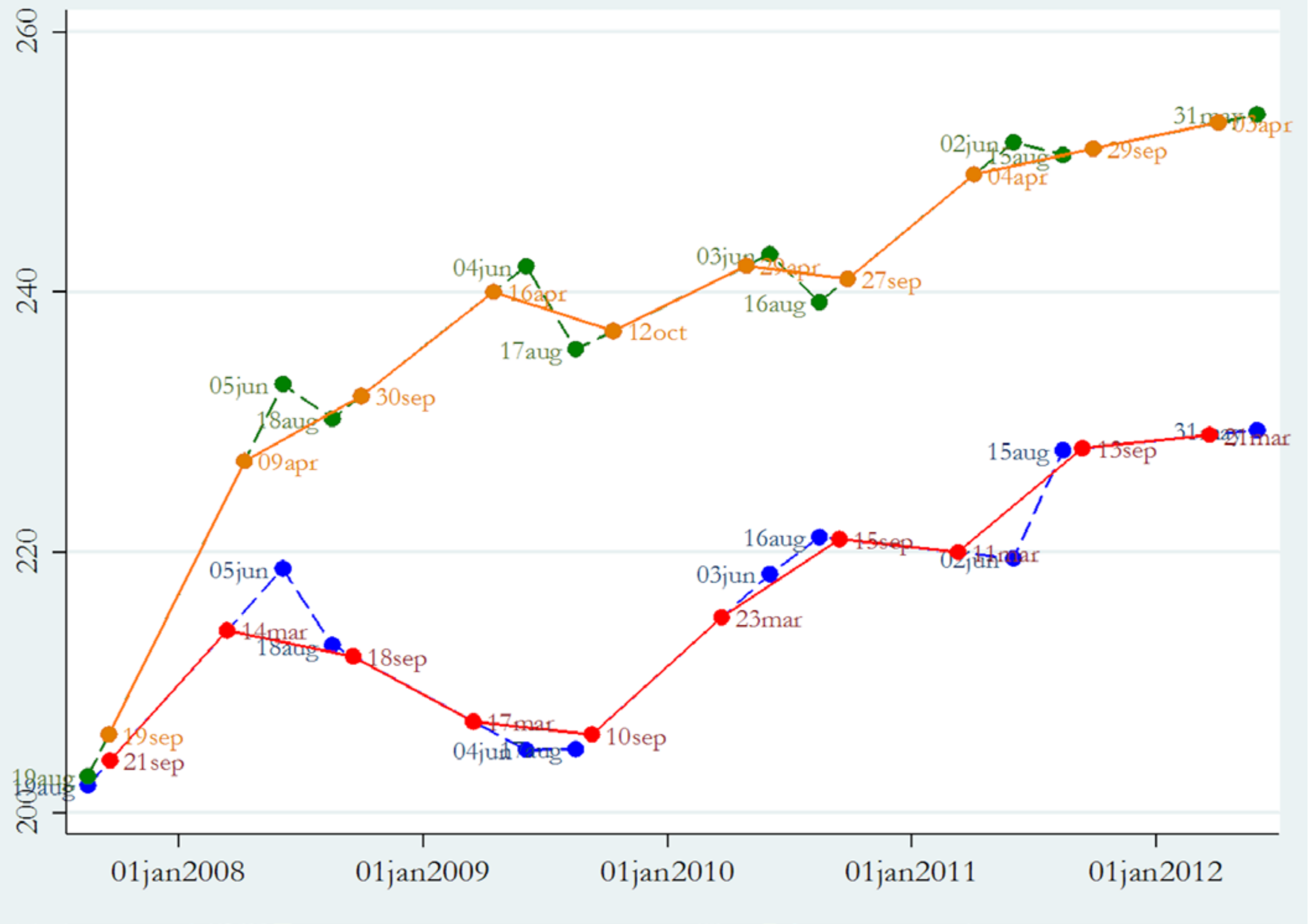
Authors	**Year of Publish**	**Dataset**	**Data Yrs**	**Years Since Data Collected**	**Summers After Grades…**	**# of Students**	**# of Schools**	**# of Districts**	**# of States**	**Geography**	**Unconditional Var(SLL)?**	**W SLL Gaps?**	**Overall SES SLL Gaps?**
Current Study	***--***	***NWEA***[5]	***2008-2016***	***3 years prior***	***1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th***	***18 million***	***~32,000***	***7500***	***50***	***Spread across all US states***	***Yes***	***No***	***No***
Heyns	1978	unnamed	1970 -1972	***47 years prior***	5th	2,978	101	1	1	Atlanta	No	Yes	Yes
Allinder, Fuchs, Fuchs, Hamlett	1992	unnamed	not found	***27 years prior at least***	2nd, 3rd, 4th	275	2	1	1	2 rural schools in midwest state	Yes (aggregated across grades)	No	No
Entwisle, Alexander	1992	BSS[1]	Fall 1982 - Fall 1984	***35 years prior***	1st, 2nd	542	20 (max)	1	1	Baltimore	No	Yes+	Yes+
Entwisle, Alexander	1994	BSS	Fall 1982 - Fall 1984	***35 years prior***	1st, 2nd	539 (max)	20 (max)	1	1	Baltimore	No	Yes+	Yes+
Alexander, Entwisle, Olson	2001	BSS	Fall 1982- Spr 1987	***32 years prior***	1st, 2nd, 3rd, 4th	678	20 (max)	1	1	Baltimore	No	Yes	Yes
Burkam, Ready, Lee, LoGerfo	2004	ECLS-K:99[2]	Fall 1998 - Fall 1999	***20 years prior***	K	3,664	~600*	~230*	~30*	nationally representative	No	No	Yes
Downey, von Hippel, Broh	2004	ECLS-K:99	Fall 1998 - Fall 1999	***20 years prior***	K	~5,000 w/ summer data*	992	~230*	~30*	nationally representative	Yes**	Yes	Yes
Borman, Benson, Overman	2005	"Teach Baltimore"	1999-2000	***19 years prior***	K (2 cohorts)	303	10	1	1	Baltimore high-povery schools	No	Yes	Yes

Figure 1 (cont'd). Compare Studies: Datasets, Data Years, Grades Included, Number of Students, Location (page 2 of 2)

Authors	Year of Publish	Dataset	Data Yrs	Years Since Data Collected	Summers After Grades…	# of Students	# of Schools	# of Districts	# of States	Geography	Unconditional Var(SLL)?	Overall B-W SLL Gaps?	Overall SES SLL Gaps?
Current Study	***--***	***NWEA*** [5]	***2008-2016***	***3 years prior***	***1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th***	***18 million***	***~32,000***	***7500***	***50***	***Spread across all US states***	***Yes***	***No***	***No***
Alexander, Entwisle, Olson	2007	BSS	Fall 1982 - 1999	***20 years prior***	1st, 2nd, 3rd, 4th	326	20 (max)	~230*	~30*	Baltimore	No	No	Yes
Benson, Borman	2010	ECLS-K:99 + Census	Fall 1998 - Fall 1999	***20 years prior***	K	4,180	290	~230*	~30*	nationally representative	No	Yes	Yes
Gershenson	2013	APSCC [3] / ATUS [4]	1989–1990 / 2003-2010	***29 years prior /*** ***9 years prior***	n/a	628 / 23,348	N/A	N/A	1 / 50	California / US	No	No	No (gaps in summer time use by SES)
Quinn	2014	ECLS-K:99	Fall 1998 - Fall 1999	***20 years prior***	K	3,043	~600*	~230*	~30*	nationally representative	No	Yes	No
Quinn, Cooc, McIntyre, Gomez	2016	ECLS-K:11	Fall 2010 - Spr 2013	***6 years prior***	K, 1st	~3,750 w/ summer data	~640*	~210*	~20*	nationally representative	No	Yes	Yes
Quinn & Le	2018	ECLS-K:99 / ECLS-K:11	Fall 98- Fall 99 / Fall 10- Spr 13	***20 years prior /*** ***6 years prior***	K / K, 1	~5,000 w/ summer data*	~600 / ~640*	~230 / ~210*	~30 / ~20*	nationally representative	No	Yes	Yes
von Hippel, Workman, Downey	2018	ECLS-K:99 / ECLS-K:11	Fall 98- Fall 99 / Fall 10- Spr 13	***20 years prior /*** ***6 years prior***	K / K, 1	~5,000 w/ summer data*	~600 / ~640*	~230 / ~210*	~30 / ~20*	nationally representative	Yes	Yes	Yes
von Hippel, Hamrock	2019	BSS / ECLS-K:99 / NWEA GRD [5]	1982-1990 / 1998-2007 / 2008-2010	***19 years prior /*** ***12 year prior /*** ***9 years prior***	1st - 4th / K / K -7 (different cohs)	825 / ~5,000* / 177,549	20 / ~600* / 419	1 / ~230* / 25	1 ~30* / 14	Baltimore / nationally rep / 14 states	No	Yes	Yes

*1 BSS (Beginning School Study); 2 ECLS-K (Early Childhood Longitudinal Study: Kindergarten Class of 1999 or 2011); 3 APSCC (Activity Pattern Survey of California Children; 4 ATUS (American Time Use Study. (time-diary surveys); 5 Growth Research Database from Northwest Evaluation Association -- subset of 25 school districts in 2008-2010. 6 Full Northwest Evaluation Association dataset-- current study. * These papers use ECLS-K:99 or ECLS-K:11 to study SLL, which can only be calculated for a subsample of ~30% of students. The student, school, district, and state N's for this subsample are not consistently reported in these papers, however we include approximate sample sizes from our direct examination of ECLS-K:1999 or ECLS-K:2011 public use datasets. N's are intended to be approximate upper bounds. **This 2004 article also examines unconditional variance, but in the authors' updated analysis, von Hippel, Workman, and Downey (2018) argue that the 2004 findings may have been affected by measurement artifacts. + These gaps are not presented by themselves, but only presented crossed with another demographic, such as SES, race/ethnicity, or school segregation status*

Figure 2. Illustration of the Timeline for Observed and Projected RIT Test Scores



FN: Student 1: Observed scores in orange, projected scores in green. Student 2: Observed scores in red, projected scores in blue.

Figure 3. ELA and Math: Estimated Mean School-year Gains and Summer Losses

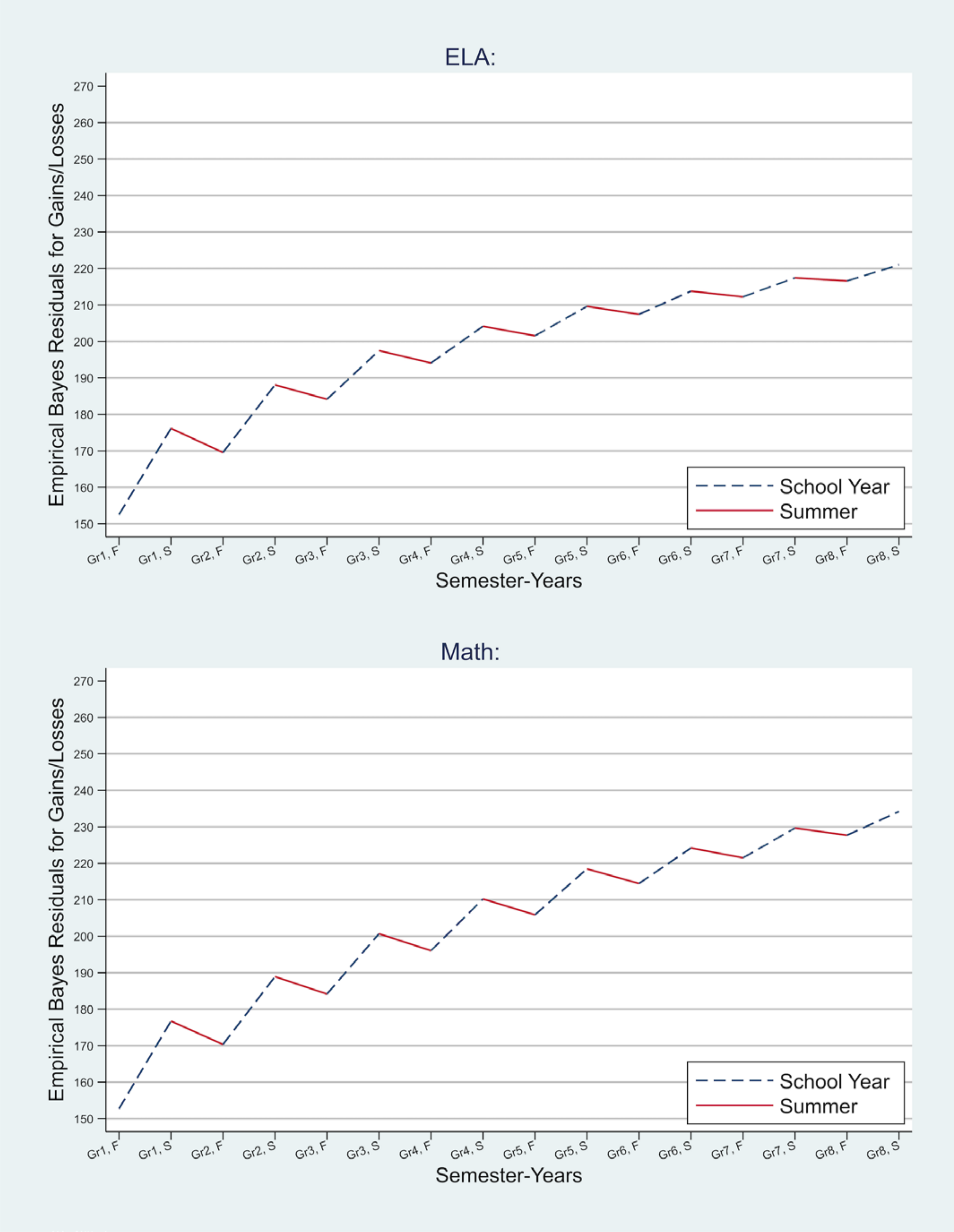
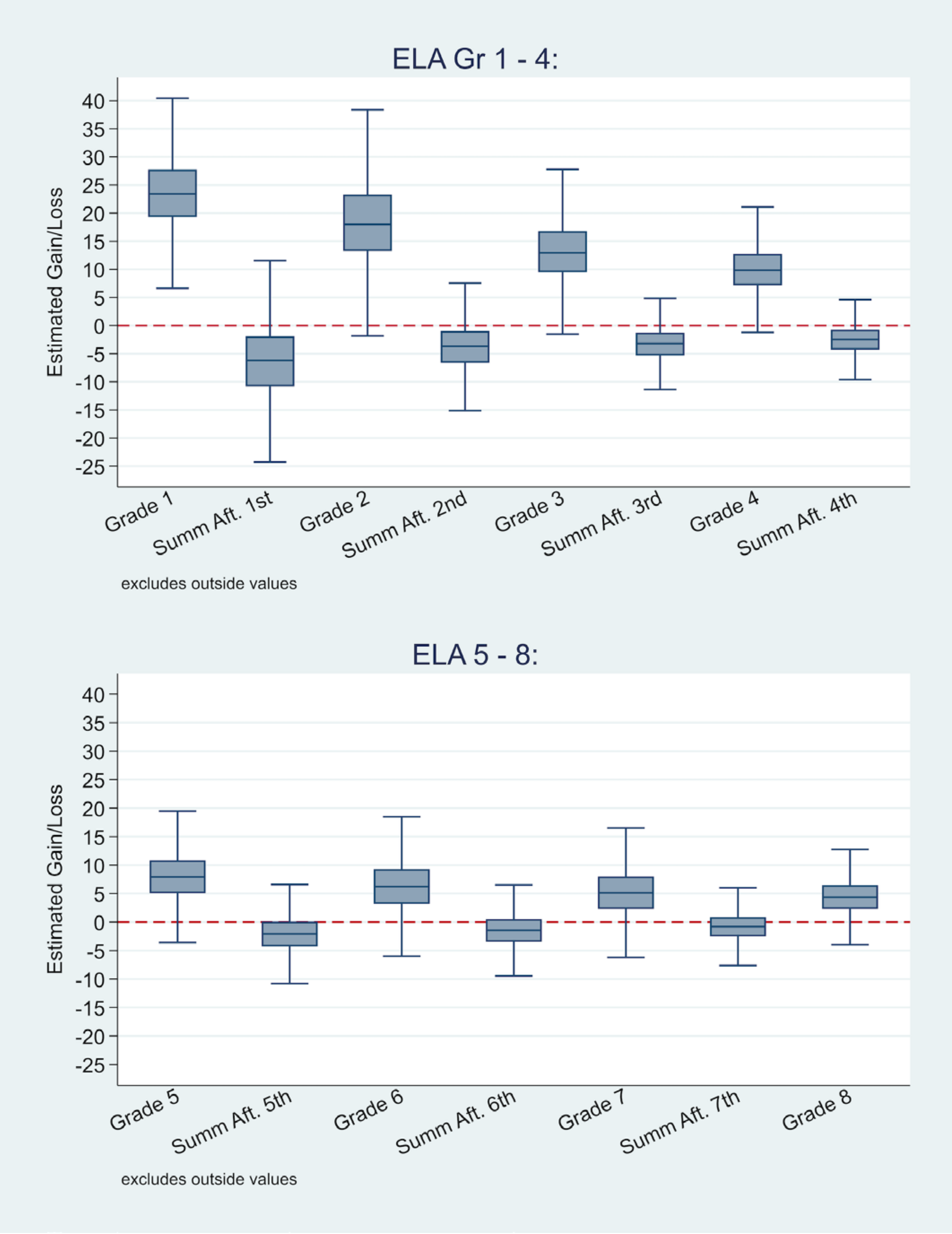


Figure 4. ELA: Boxplot of Students' Empirical Bayes Estimated Gains/Losses, across Grades



ELA Gr 1 - 4:

excludes outside values

ELA 5 - 8:

excludes outside values

Figure 5. Math: Boxplot of Students' Empirical Bayes Estimated Gains/Losses, across Grades

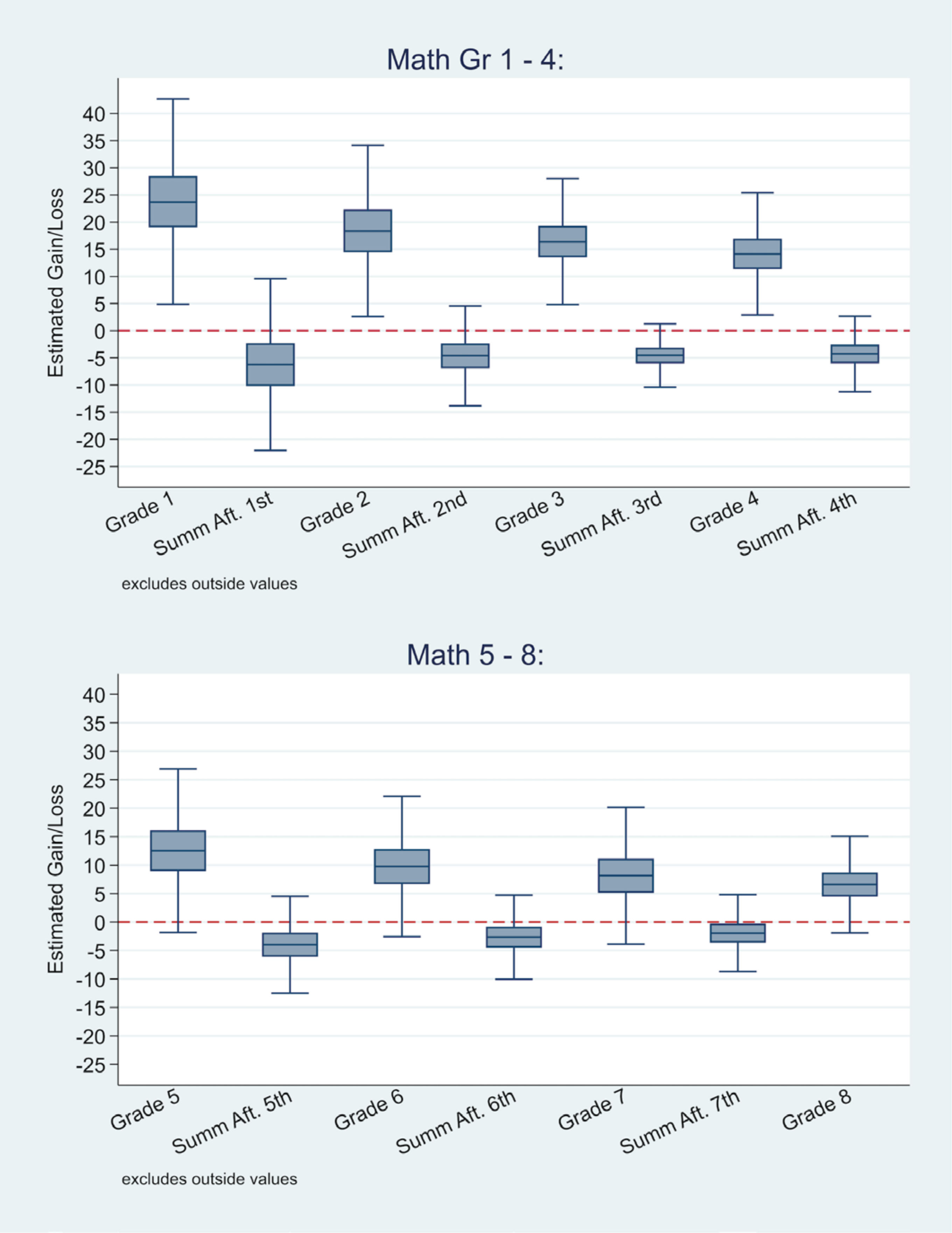


Figure 6. Math and ELA: Assume Equal Learning in School, Three Levels of Summer Gains/Losses

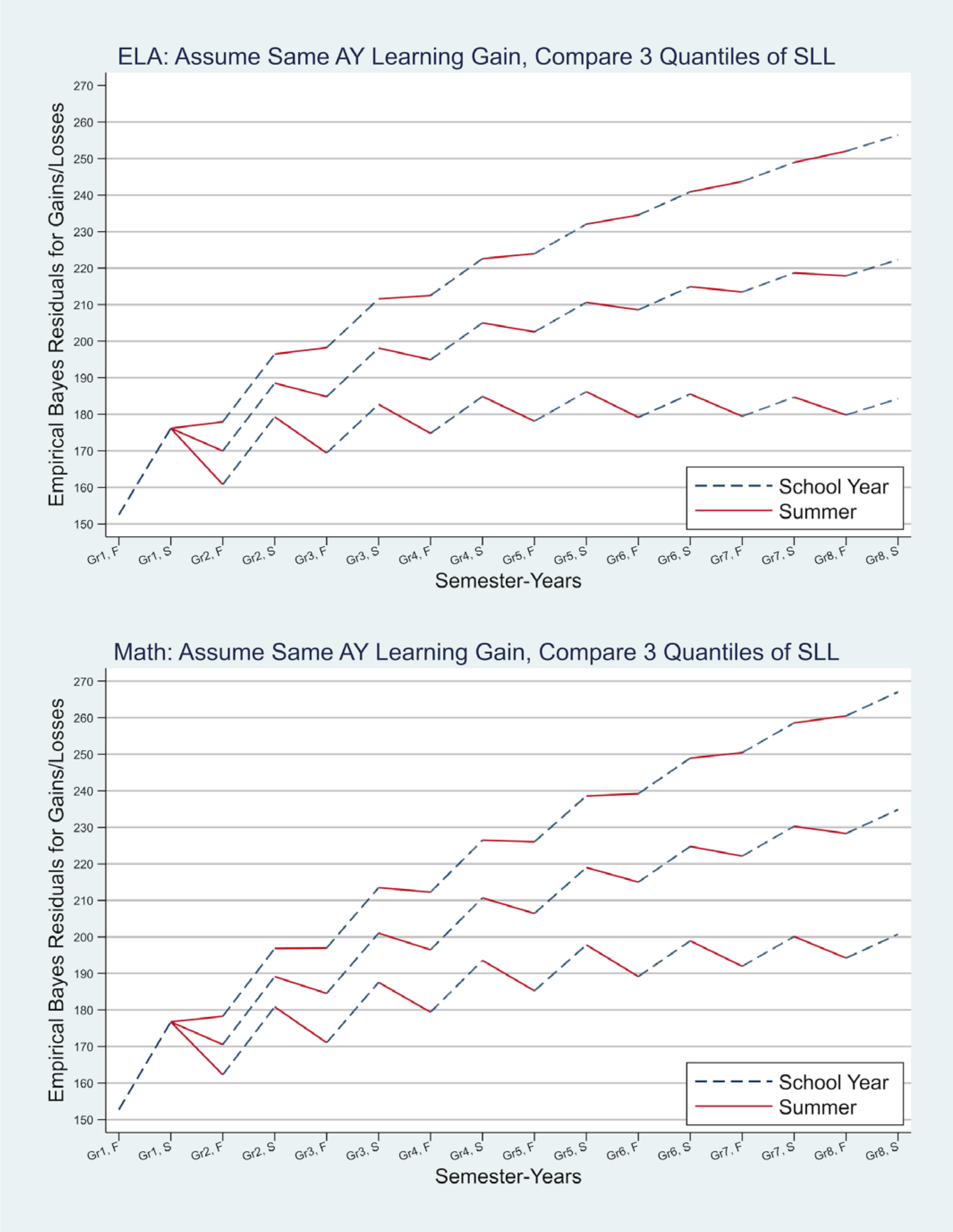
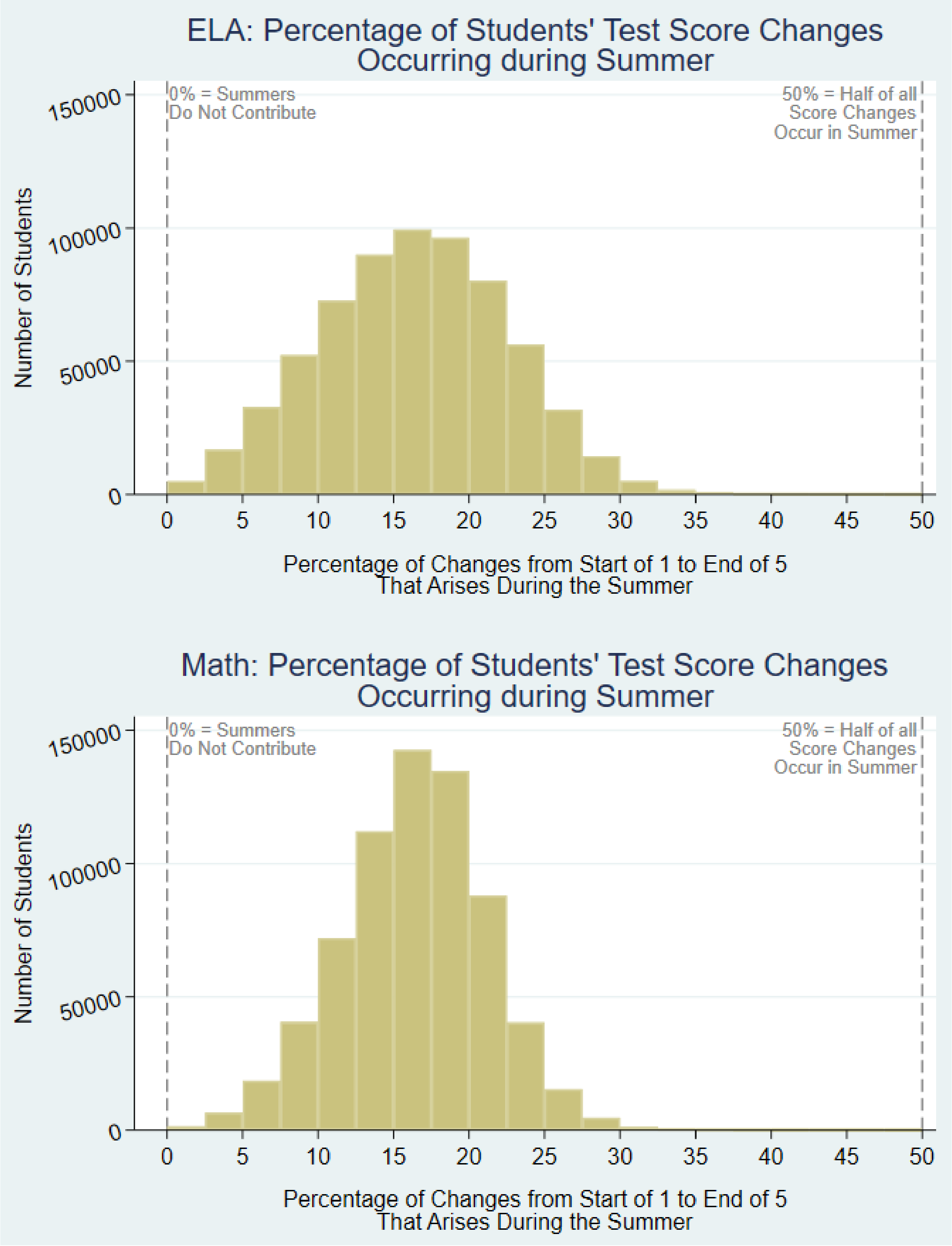


Figure 7. ELA and Math: Proportion of Students' Test Score Fluctuations Occurring in Summers



ELA: Percentage of Students' Test Score Changes Occurring during Summer

Math: Percentage of Students' Test Score Changes Occurring during Summer

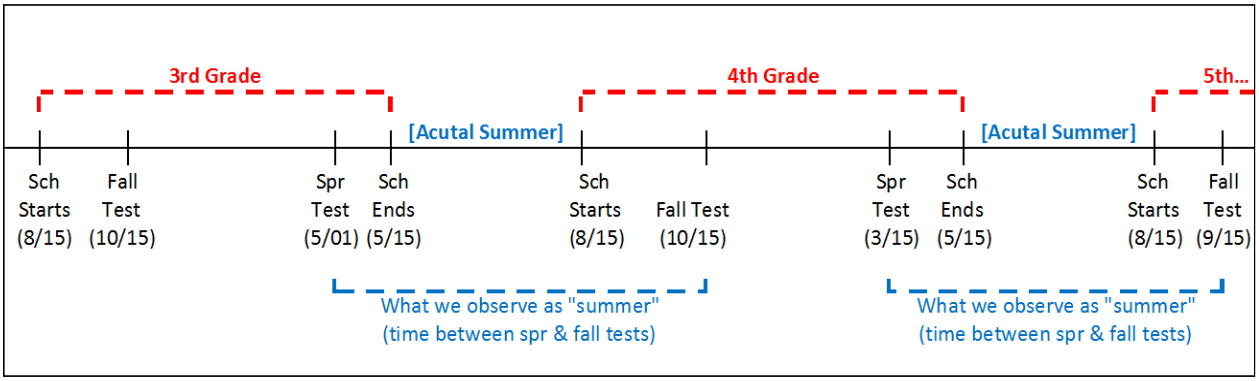
Online Appendix A. School Calendar Dates and Projected RIT Scores

Approach to Calendar Data Collection

One unique aspect of the current project was to collect, clean, and incorporate a new source of information about school-years into both the current analyses, as well as to share the information back with our research partner NWEA to improve their own internal analyses. We collected longitudinal information on school calendars at the district level for all districts in a set of eleven states that have the largest percentage of students with MAP scores. In fact, 44.4% of all student-year observations from the NWEA data come from this subset of states.

In Figure 1, we show a hypothetical timeline for a given student's test-taking from 3rd through 5th grade. The Figure illustrates that students do not take MAP tests *exactly* on the first and last day of school—in fact, students often take these tests three to six weeks before/after the school-year starts or ends. As a result, some of the time between the spring and fall administrations of the test is actually spent in school. However, we do not observe school-year start or end dates, leaving us with a distorted sense of how long students spend without the structure of the school-year—the very time when we suspect learning rates may slow. Without knowledge of school-year calendars, we would misattribute some of the learning that takes place during the school-year to the summer period, potentially masking some of the actual variation in the summer period. We therefore obtained the school-year calendar information through original data collection.

<u>**Figure A1. Illustration of the Need for School Calendar Data**</u>



The scope of this data collection task varied considerably, and it depended largely on whether each state has adopted a statewide policy on school-year start- and end-dates, or whether state departments of education kept this information in existing data files. For example, the process for South Carolina was relatively simple because, beginning in August of 2007, South Carolina adopted new statewide legislation that specified consistent school start and end dates. We have found online a document that reported each of South Carolina school districts' calendars from 2010-11 through 2015-16. We examined the extent to which school districts actually used the uniform start and end dates mandated by the legislation (district level calendars are no longer available prior to 2010-11). In the years of district-level calendar that we have, it appears that the vast majority of South Carolina districts uses the same school-year start and end dates that is described in the legislation: School typically starts on the third Monday of August, and the last day of school falls on the first Thursday of June.

In the eleven other states in which we conducted data collection, there is no statewide legislation that specifies district start and end dates. To gather the data in other states and years, we worked with a team of undergraduate and graduate student research assistants in efforts to collect complete records on school district calendars across our twelve-state sample. The first step was to exploit all online resources to find existing records from state- and district-level education departments. We also used an internet archive website (https://archive.org/web/) to search for this information that had potentially been archived in prior years. In cases where such documentation could not be found, research assistants also examined news sources archived online that document district-wide school calendars. We found that newspapers often run stories about the school-year timeline. Finally, once all indirect methods of obtaining school calendar records have been exhausted, research assistants contacted appropriate district or state personnel directly to request the information.

Altogether, we proposed collecting school-year start and end dates in 3,119 unique districts across eleven states and nine school-years, for a total of about 28,000 district-years. We collected 23,223 school-year start dates and 20,807 school-year end dates. We therefore found about 77% of the district-year calendar dates we sought to find. In Table 1, we present the percentage of districts in each state and in each year for which we have collected school-year start dates. In green we highlight cells that have over 90% coverage, and in red we highlight cells that have less than 55% coverage.

Table A1. District Coverage (Percentage), by State and Year

State	2008	2009	2010	2011	2012	2013	2014	2015	2016
CO	23.6	23.6	23.6	25.1	24.6	25.6	26.1	86.2	77.6
IA	93.1	96.4	96.4	96.4	96.4	96.4	95.8	93.6	93.1
IN	99.1	99.1	99.1	99.2	99.2	98.7	98.7	98.7	98.7
KS	95.6	95.9	96.2	96.5	96.3	96.0	96.3	96.3	94.9
KY	10.9	9.8	100.0	100.0	100.0	100.0	99.4	99.4	99.4
MN	99.4	99.4	99.6	99.6	99.6	99.6	98.5	97.4	54.0
NH	97.7	97.7	97.7	97.7	97.7	97.7	97.2	97.2	96.7
SC	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
WA	31.9	43.4	61.1	65.5	65.2	63.9	75.7	91.2	96.3
WI	99.3	98.2	98.6	99.5	99.6	99.6	99.6	99.6	0.0
WV	100.0	100.0	100.0	98.2	100.0	100.0	100.0	100.0	0.0

In later years, NWEA also began to collect some school-year start and end dates. We combined our original data collection with theirs. Across the entire NWEA dataset in all states and years, these efforts allowed us to collect actual calendar start/end dates for 50.3% of the observed school-years. We refer to these as the "actual calendar dates", because we also opt to extrapolate calendar dates for all districts in which they are missing.

Using Actual Calendar Dates to Extrapolate Missing Calendar Dates

In order to project scores for students in districts for which we were unable to recover actual calendar dates, we chose to impute approximate school calendar dates under the basic assumption that, while there is some variation in when public school districts start and end the school-year, it is not large. For example, in the subset of districts for which NWEA collected *school* level calendar start dates, we observe that the standard deviation of start dates across schools in the same district and same year is 8.2 days (8.1 days for end dates). Looking across all the districts in a given state in the same year, the standard deviation of start dates is 6.3 days (8.2 for end dates).

We therefore extrapolated dates privileging the following, ordered decision rules: (1) If we have actual school calendar data, use that. (2) For schools in a district-year with some school calendar data, use the mean of the start/end dates in the district-year. (3) For a district that has calendar data in some years but not others, use the district's own mean start/end dates across years. (4) For districts still missing start/end calendar dates, use the state's mean dates in the given year. (5) For districts in states that have no calendar data in a given year, use the state's mean calendar dates across all years. This covered all observations in the dataset.

We tested this approach with the following exercise: We limited the sample to district-years with actual school calendar dates, hid a random sample of 25% of the actual dates as if they were missing, and then used the procedure above to produce extrapolated dates for that 25%. We then compare the extrapolated dates to the actual dates to assess the extrapolation process. On average, the process produced an extrapolated start date 3.8 days off from the actual start date and 6.6 days off from the actual end dates, providing additional confidence in the extrapolation process.

Projecting RIT Scores to First and Last Day of School

We leverage the calendar data described above to project scores for individual students to what they might have been on the first and last day of school. We calculate the average daily learning rate between each student's fall and spring test administrations by dividing the change in score by the number of days between the two tests (Quinn, 2014). Extant research finds that students' within school-year achievement growth is approximately linear (Fitzpatrick, Grissmer, & Hastedt, 2011). We then calculate both the number of school days between the start of the school-year and each student's fall NWEA test, as well as the number of school days between each student's spring NWEA and the end of the school-year. On average, students take the fall test about 26 days after the first day of school and the spring test 39 days before the last day of school.

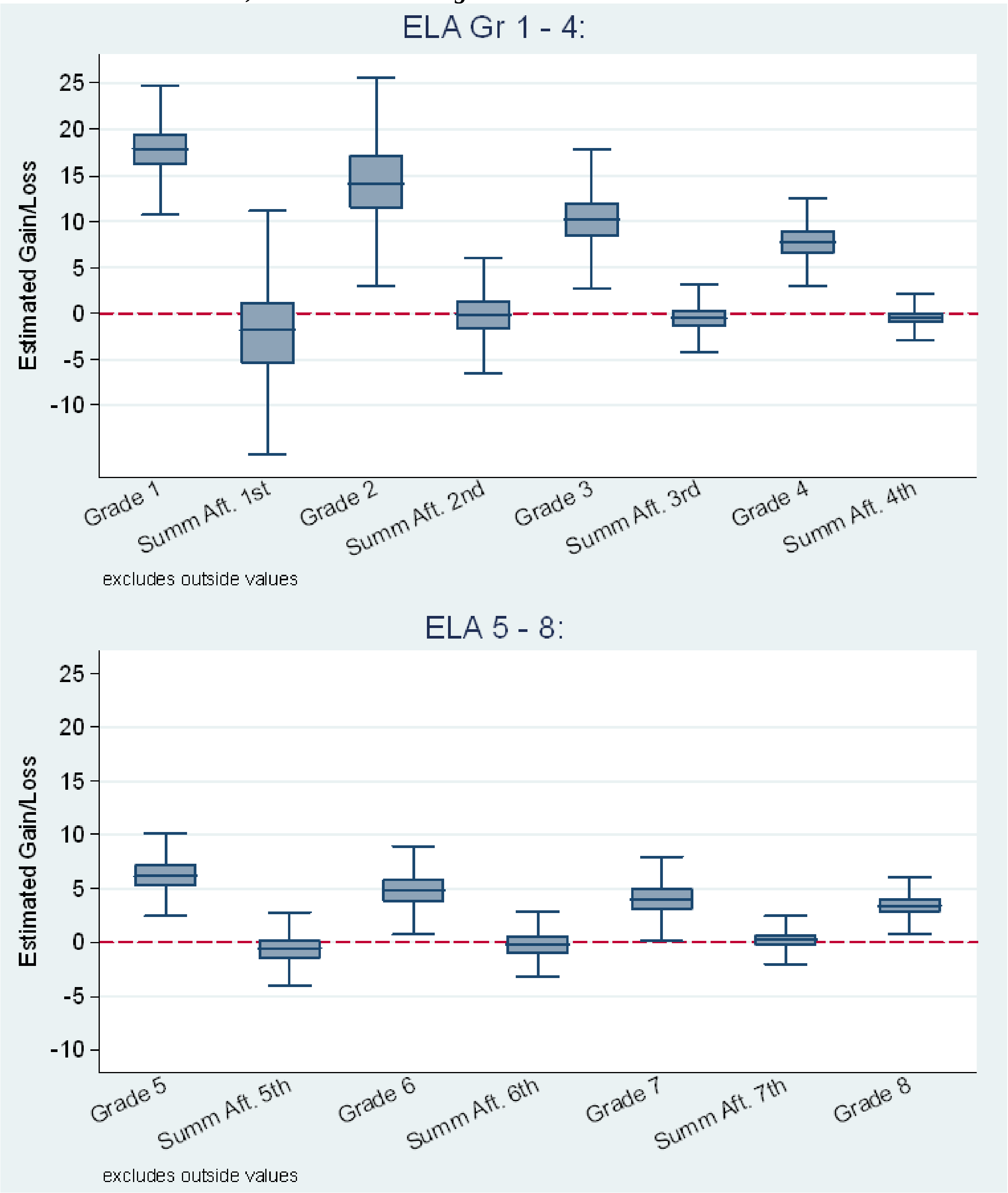
To project scores to the start of the school-year, we subtract from the student's observed fall score his or her individual daily learning rate multiplied by the number of days between the first day of school and the date of the test. We follow the same procedure for projecting scores to the last day of the school-year. The correlation between fall observed and projected scores in ELA is 0.996, with an RMSE of 2.3 points. The correlation between spring observed and projected scores in ELA is 0.992, with an RMSE of 2.8 points.

Appendix A. References

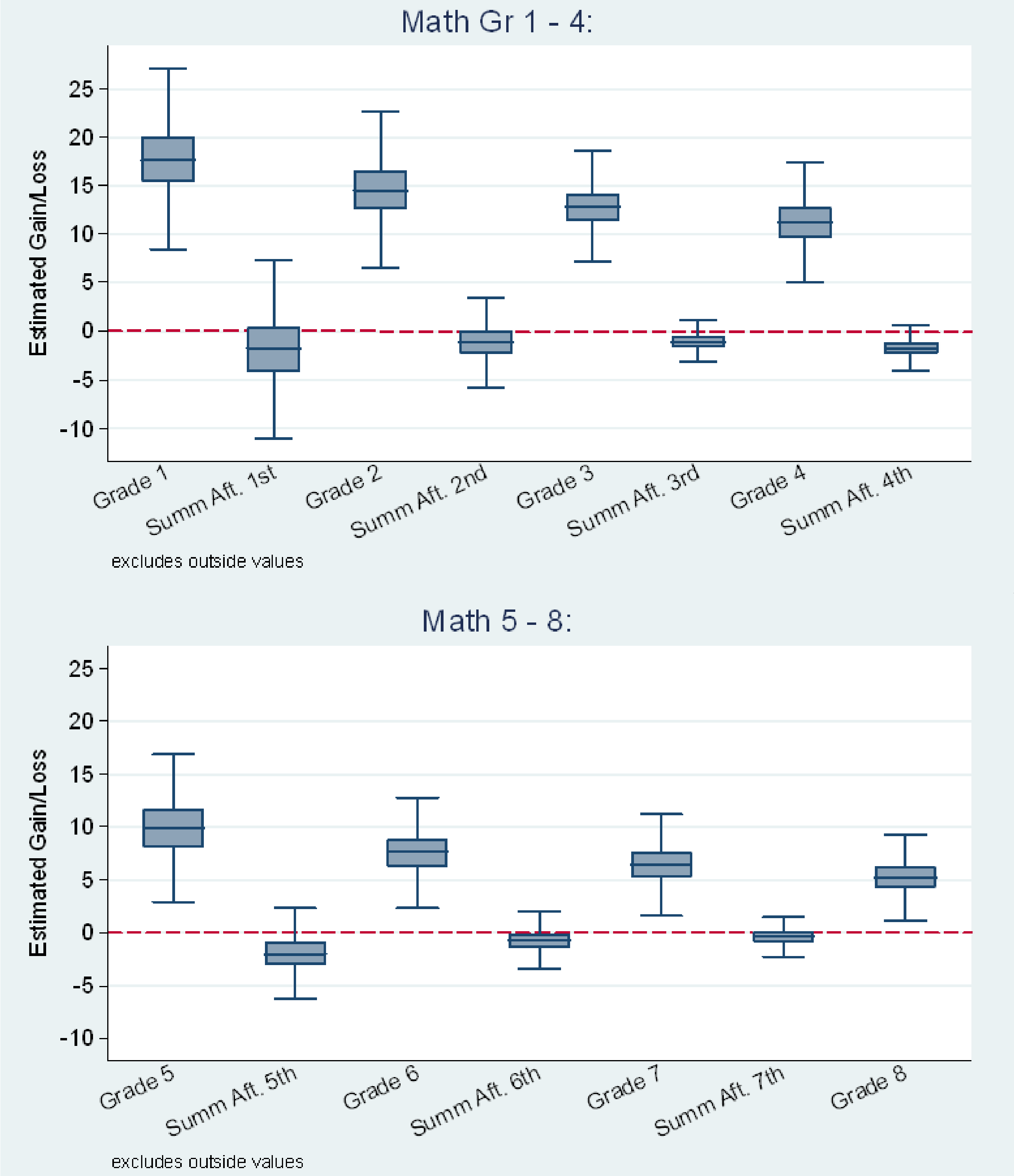
Fitzpatrick, M. D., Grissmer, D., & Hastedt, S. (2011). What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review*, 30(2), 269-279.

Quinn, D. M. (2014). Black-White Summer Learning Gaps: Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis*. doi:10.3102/0162373714534522

Online Appendix A1. Replicated Narrative Figure 4. ELA: Boxplot of Students' Empirical Bayes Estimated Gains/Losses, across Grades using Observed Scores



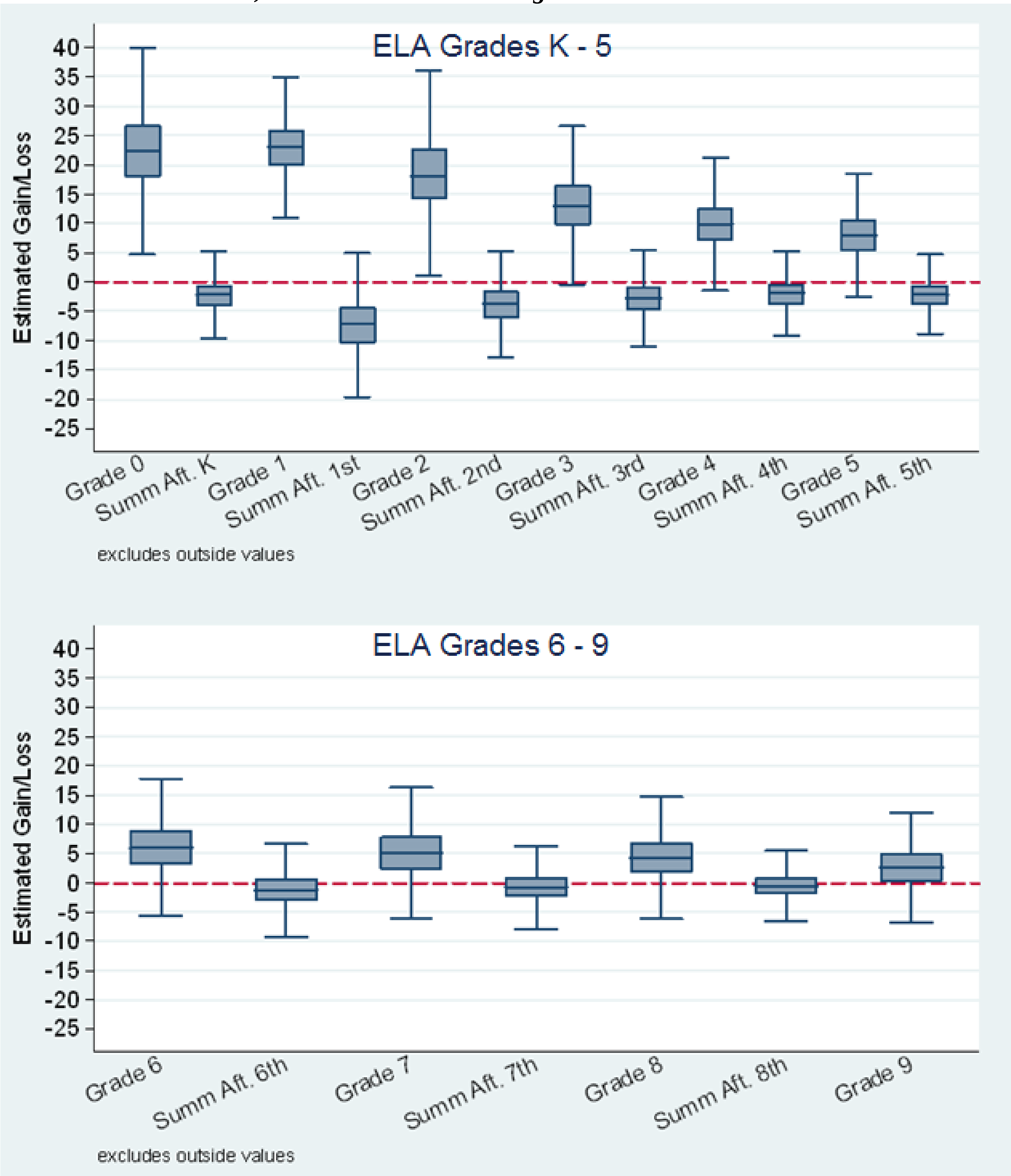
Online Appendix A2. Replicated Narrative Figure 5. Math: Boxplot of Students' Empirical Bayes Estimated Gains/Losses, across Grades using Observed Scores


Math Gr 1 - 4:

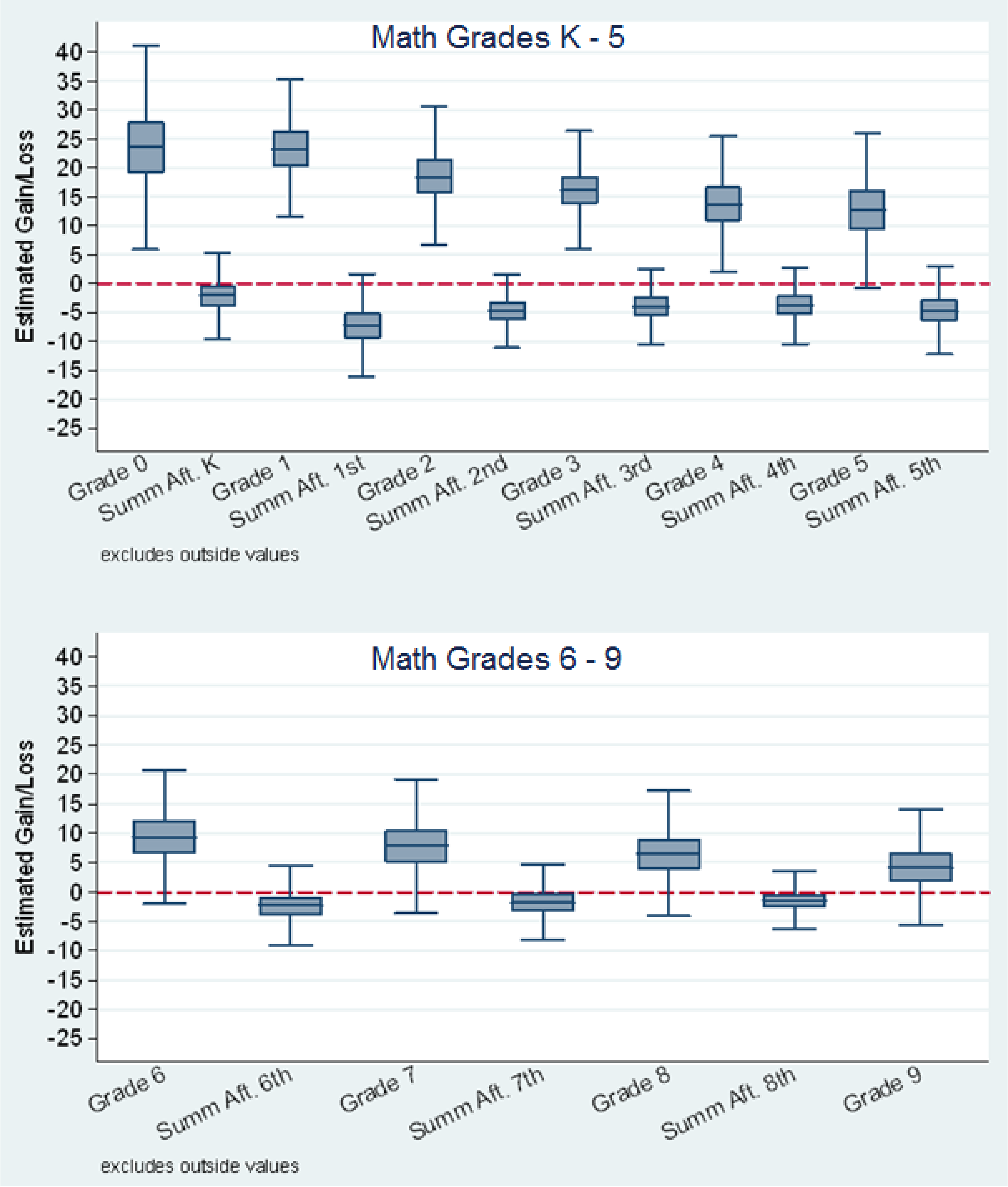
Math 5 - 8:

Online Appendix B. Replicate Figure 4 & 5 with 3-Grade Increment, Grade K (0) – 9

Online Appendix B1. Replicated Narrative Figure 4. ELA: Boxplot of Students' Empirical Bayes Estimated Gains/Losses, across Grades K – 9 using a 3-Grade Increment



Online Appendix B2. Replicated Narrative Figure 5. Math: Boxplot of Students' Empirical Bayes Estimated Gains/Losses, across Grades K – 9 using a 3-Grade Increment



Endnotes

[1] We use the term disparities broadly to refer to any large, potentially systematic variability across students. While some use the term disparities to refer to group mean differences (e.g., by race/ethnicity or SES), we use disparate or disparities synonymously with spread or variability across students. We use gaps only to for group mean differences.

[2] Race and SES gaps have been the main focus of prior SLL research.(Alexander, Entwisle, & Olson, 2001; Alexander, Entwisle, & Olson, 2007; Benson & Borman, 2010; Borman, Benson, & Overman, 2005; Burkam, Ready, Lee, & LoGerfo, 2004; Downey, von Hippel, & Broh, 2004; Entwisle & Alexander, 1992, 1994; Heyns, 1978; Quinn, 2014; Quinn, Cooc, McIntyre, & Gomez, 2016; von Hippel & Hamrock, 2019; von Hippel, Workman, & Downey, 2018)

[3] Given the importance of assessing the role of SLL on the development of racial/ethnic and SES achievement gaps, some of their methodological nuances (Quinn, 2014), and the large amount of variability that remains unexplained by these demographics, exploring race and/or SES gaps in SLL deserves its own separate and full investigation. The goal of the current paper is to update the existing knowledge base about overall 1st through 8th grade school-year learning gains and subsequent summer loss patterns, document the degree of variability in those patterns, and characterize the extent to which end-of-school achievement disparities arise during summers.

[4] This and many SLL studies specifically examine the summer after kindergarten.

[5] Strictly speaking, most studies actually show that, on average, students do not lose ground during the summer, but instead either gain less in the summer than in the school-year (learning rate slows) or have no gains in the summer.

[6] A series of studies followed that examined SLL in specific locations (e.g., Allinder, Fuchs, Fuchs, & Hamlett (1992) in 2 rural schools around 1990; Borman, Benson, & Overman (2005) with about 300 students in Baltimore high poverty schools; Skibbe, Grimm, Bowles, & Morrison (2012) with about 380 students in 1 suburban Midwest town). That said, it has been unclear whether the results from those early studies would either generalize outside of their local contexts or to a vastly different educational landscape up to forty years later.

[7] In ECLS-K:99, the target number of children sampled at any one school was 24, and on average 5.8 students were sampled per classroom (based on our analysis of publicly-available ECLS-K:99 data, but see also similar reported classroom sample sizes in Gershenson and Hayes (2018)). However, because only one third of the K students were sampled for fall testing in grade 1, on average, only 1.5 students per K classroom (3.2 students per K school) possess both the K spring *and* 1st fall scores needed to estimate their SLL in the summer after K. About 18% of K classrooms (and schools) in ECLS-K have more than 3 students with SLL estimates.

[8] We provide a brief summary of their findings with respect to race and SES gaps, using scores that were not standardized by subject-semester-grade, the preferred theta scale from ECLS-K:99, and the most comparable grade ranges: With respect to Black-White race gaps, the authors find—across the three datasets—that gaps grow across grades (with the exception of an aberrant finding from the BSS of 556% shrinkage of the Black-White ELA gap across grades), though that growth is more moderate in the ECLS-K:99 and GRD datasets. In both the BSS and the ECLS-K:99 datasets (preferred theta scores), the authors find that there is no significant difference in how fast the Black-White gap grows in the summer vs. the school-year. However, in the more recent GRD dataset, it appears these gaps grow more during the school-year. With respect to SES gaps, von Hippel and Hamrock (2019) find that, while student-level SES gaps appear to grow across grades in the BSS, they appear to shrink in ECLS-K:99. Gaps in low- versus high-poverty schools seem to grow using BSS data and GRD data (though to a smaller degree), but shrink using ECLS-K:99 data. Both ECLS-K and BSS datasets show that student SES gaps grow faster in the summer, as opposed to the school-year, and all three datasets indicate that low- versus high-poverty school gaps grow faster in the summer (with the exception of math results using the GRD).

[9] Here, we highlight the comparison between the ECLS-K IRT-based scale scores (which estimate the number of items a child would correctly answer and is not designed for comparison over time) in their original metric (i.e., not standardized by subject-semester-year) versus that same dataset's IRT-based theta scores in their original metric.

[10] Rambo-Hernandez and McCoach (2015) use a sample that follows a 2006-2009 cohort of about 118 thousand grade 3 students as they move through grade 6.

[11] It is also administered in the winter by some districts; however, the winter data are not available in the current dataset.

[12] We also conduct analyses presented in Table 2, Table 3, Figure 4, and Figure 5 using only the set of district-years with actual school-year start/end dates (rather than extrapolated dates). Results are quite similar and available upon request.

[13] To contextualize the RMSE, NWEA reports the achievement status norm for ELA is about 161 in fall of grade 1 and about 217 in fall of grade 8 (Thum & Hauser, 2015).

[14] Because the summer learning rate is estimated off of just two points—the first and last day of school—the slope between those points is quite sensitive to even minor adjustments. Note that the method we describe assumes that students learn just as much on days in May as they do in, say, February. While there is some evidence that learning rates are relatively linearly within the school-year (Fitzpatrick, Grissmer, & Hastedt, 2011; von Hippel & Hamrock, 2019), there are also reasons to question this assumption, especially given anecdotal reports that the intensity of school activities slows after spring standardized test are given.

[15] The analytic samples in this paper are first limited to NWEA students observed in grades 1 through 8, hence the large drop in sample size between the full NWEA sample and the Appendix B analytic sample.

[16] See Raudenbush and Bryk (2002), pp. 55-56 for a more complete description of the Bayesian approach to estimation of the variances and covariances. For a discussion of how the observed variability in OLS estimates compares to the empirical Bayes estimate of the variability, see pg. 88.

[17] We include 5th grade in both panels to informally check how similar 5th grade estimates are across the models.

[18] For example, $schyr6_{ti}$ takes a value of 1 at the end of 6th grade (i.e., grade 6 spring test score) and remains a 1 for all observations thereafter. And $sumaf6_{ti}$ takes a value of 1 at the end of the summer after 6th grade (i.e., grade 7 fall test score) and remains a 1 for all observations after.

[19] For example, Downey et al. (2004) code time variables so that the relevant coefficients thereon represent a linear learning *rate per month* between the first and last day of school (or first and last day of summer). In contrast, we have chosen to code time dummies so that the relevant coefficients capture the *total gain* from the first to last day of a given school year (or the *total gain/loss* from the first to last day of summer). As a concrete example, suppose that a hypothetical student gained a total of 9 test score points during the school year but lost 3 of those test score points during the subsequent summer. Under the coding scheme used by Downey et al. (2004), the coefficients would be expressed in points per month: +1 in the school year versus -1 in the summer. Under the coding scheme used herein, the coefficients would be expressed in total gains/losses—that is, +9 in the school-year compared to -1 in the summer. This example illustrates how using learning rates makes it more difficult to appreciate what proportion of the school year gain was lost during the summer. In addition, presenting the estimates as a monthly learning rate may imply to some readers that we have data on what happened on a *monthly* basis and that the function is, indeed, linear.

[20] The parameters are presented with a focus on their substantive meaning in the Results section, but for those interested in a more formal roadmap between research questions and parameters: For RQ1 concerning mean gains/losses, we focus on the β coefficients. For RQ2 concerning student-level variation in gains/losses, we interpret the $\boldsymbol{\tau}$ variance parameters from the diagonal of the covariance matrix. For RQ3 concerning whether the same students tend to lose ground summer after summer, we present the off-diagonal elements of the covariance matrix corresponding to π_{2i} and π_{4i}, as correlations (take, for example, the relationship between losses in the summers after 6th vs. 7th grade; for this example, the covariance is $\boldsymbol{\tau_{2,4}}$). For RQ 4, we make use of the student-level Bayes shrunken residuals.

[21] Burkam et al.'s (2004) SLL analysis of ECLS-K:1999 data shows that, taken together, students' gender, racial, and socio-economic demographics in conjunction with detailed information from parent surveys about children's home and summer activities only accounts for about 13% of the variance in learning gains in the summer after K.

[22] Returning briefly to Equation (1) for a concrete example, consider the covariance of the π_{2i}'s (student i's estimated change in the summer after 6th grade) with the π_{4i}'s (in summer after 7th grade). That covariance ($\tau_{4,6}$) from the covariance matrix captures the extent to which students who lose ground in one summer tend

to be the same ones who lose ground in the next summer. Like the variances presented earlier, these estimated covariances are more conservative than simply taking the standard deviation of student-level gain/loss scores (Raudenbush & Bryk, 2002). We present the covariance as correlations for ease of interpretation.

[23] The one exception to the otherwise uniformly-negative correlations in the lower panel of Table 4 (ELA) is the near-zero correlation of +0.01 between grade 1 school-year gains and gains/losses in the summer after grade 4.

[24] In von Hippel et al. (2018), 11 of these 12 reported correlations are negative and between -0.55 to -0.21, with the one exception of a modest, positive correlation ($+0.09$) between the ELA learning rates in the summer after grade 1 and the grade 2 school year.

[25] We split the distribution of student-specific, empirical Bayes shrunken summer learning gain/loss estimates into a top, middle, and bottom tercile and then calculate the mean learning gain within each of those terciles. We do this separately for residuals for each summer following a school-year between first and 8th grade.

[26] We calculate for each student the sum of all absolute fluctuations in their test scores during a panel (here, from the start of 1st to the end of 5th grade) and then calculate what percentage of those absolute value fluctuations arose during summers. For a hypothetical student who experiences no change in their scores from the start to the end of the summers (i.e., always flat slopes in the summers), this percentage would be zero. In contrast, if a hypothetical student's test score changes during the summer were always equal to the student's gain/loss during the school-year, the corresponding statistic would be 50%.

[27] Looking across the full study sample, about 75% of *all* summer period changes were negative (as opposed to gains or no change). If summer loss events were truly independent, the probability of 5 consecutive summer losses is 0.75 raised to the 5th power, which equals about 0.24.

[28] For instance, von Hippel, Workman, Downey (2018) find that students exhibit slightly greater SLL in reading than in math in the summer after K, but equal losses across subjects in the summer after grade 1. Descriptive results from Quinn, Cooc, McIntyre, & Gomez (2016) suggest that students gained very similar amounts in math and reading during summers, but perhaps gained slightly more in math in the summer after K and slightly more in reading in the summer after grade 1. Downey, von Hippel, & Broh (2004) document modest mean summer losses in reading, alongside summer learning gains in math. Many of the studies since 1996 that specifically present mean SLL rates, only present these results for a single subject, preventing a cross-subject comparison (e.g., Borman, Benson, & Overman (2005); Downey, von Hippel, Hughes (2008); Benson & Borman (2010); Skibbe, Grimm, Bowles, & Morrison (2012); Rambo-Hernandez & McCoach (2015).

[29] Downey et al. (2004) also do so, but they subsequently discount those findings and update them in von Hippel et al. (2018).

[30] The authors describe the achievement measures they use from ECLS-K:99 as "IRT scores" (see Table A1 on pg. 424), and we believe these are likely the IRT-based scale scores (rather than IRT-based theta scores), which model the number of items children would have answered correctly, using summed probabilities of correct answers (Tourangeau et al., 2009). It is beyond the scope to reconcile this difference across ECLS-K studies, but this warrants further attention.