sociological science

Findings on Summer Learning Loss Often Fail to Replicate, Even in Recent Data

Joseph Workman,[a] Paul T. von Hippel,[b] Joseph Merry[c]

a) University of Missouri, Kansas City; b) University of Texas, Austin; c) Furman University

Abstract: It is widely believed that (1) children lose months of reading and math skills over summer vacation and that (2) inequality in skills grows much faster during summer than during school. Concerns have been raised about the replicability of evidence for these claims, but an impression may exist that nonreplicable findings are limited to older studies. After reviewing the 100-year history of nonreplicable results on summer learning, we compared three recent data sources (ECLS-K:2011, NWEA, and Renaissance) that tracked U.S. elementary students' skills through school years and summers in the 2010s. Most patterns did not generalize beyond a single test. Summer losses looked substantial on some tests but not on others. Score gaps—between schools and students of different income levels, ethnicities, and genders—grew on some tests but not on others. The total variance of scores grew on some tests but not on others. On tests where gaps and variance grew, they did not consistently grow faster during summer than during school. Future research should demonstrate that a summer learning pattern replicates before drawing broad conclusions about learning or inequality.

Keywords: summer learning; summer learning loss; summer setback; summer slide; replication; replicability

Parents, teachers, scholars, and advocates often voice concern about children losing academic skills over summer vacation. According to the National Summer Learning Association (NSLA 2017), "most students lose two months of mathematical skills every summer," and losses are greater among "low-income children, [who] typically lose another two to three months in reading." As a result, NSLA (2017) claims, "summer learning loss during elementary school accounts for two-thirds of the achievement gap in reading between low-income children and their middle-income peers by ninth grade."

Claims about summer learning loss and summer learning gaps do not just raise practical concerns for parents and teachers. They also raise fundamental issues about where inequality in school performance comes from and how it can be reduced. The claim that score gaps between advantaged and disadvantaged children grow fastest during summer implies that the bulk of socioeconomic inequality in educational achievement comes from outside schools rather than inside them, and not just in early childhood before school begins, but every time school lets out for summer vacation (Downey and Condron 2016; Downey, von Hippel, and Broh 2004). If this is true, then interventions that focus only on school might leave the major sources of inequality unaddressed, whereas interventions that focus on summer and the out-of-school environment might have greater potential—not just to raise average achievement but to reduce achievement gaps.

Citation: Workman, Joseph, Paul T. von Hippel, and Joseph Merry. 2023. "Findings on Summer Learning Loss Often Fail to Replicate, Even in Recent Data." Sociological Science 10: 251-285.

Received: January 6, 2023

Accepted: February 6, 2023

Published: March 30, 2023

Editor(s): Arnout van de Rijt, Richard Breen

DOI: 10.15195/v10.a8

Some scholars, however, have raised concerns about the replicability of claims about summer learning. “The recent literature on summer loss,” Quinn and Polikoff (2017) wrote, “has been mixed.” Results have been mixed even for the simple question of whether children lose skills on average over the summer. Whereas some data do suggest that most students lose months of reading and math skills every summer (e.g., Atteberry and McEachin 2021), other data show practically no summer losses on average (Quinn and Polikoff 2017; von Hippel, Workman, and Downey 2018).

Results are also mixed on NSLA’s claim that the score gaps between children of high and low socioeconomic status (SES) grow substantially during summer—or at least grow faster during summer than they grow during the school year. Whereas some data did suggest that SES gaps grow dramatically during summer and not during school (Entwisle and Alexander 1994), other data suggest that SES gaps grow no faster during summer than during school (von Hippel and Hamrock 2019). Some data, in fact, show SES gaps growing little, or even shrinking, over the whole period from kindergarten through eighth grade, leaving little room for summer learning—or school-year learning—to contribute much to the gap between high- and low-SES children at the start of ninth grade (Duncan and Magnuson 2011; Heckman and Masterov 2007; von Hippel and Hamrock 2019; von Hippel et al. 2018; Yen, Burket, and Fitzpatrick 1995b).

Summer learning findings have also been mixed for the score gap between black and white children. Although one older study reported that the black–white gap grew mainly during summer (Hayes and Grether 1983), some more recent studies reported that the black–white gap grew mainly during the school year (Downey et al. 2004; Entwisle and Alexander 1994; von Hippel et al. 2018), and one study reported that summer learning rates were different for black children attending segregated or integrated schools (Entwisle and Alexander 1994). A few recent studies reported that black children actually *gain* on white children during summer (Kuhfeld et al. 2020; von Hippel and Hamrock 2019; von Hippel et al. 2018), but this finding is hard to reconcile with evidence that black families have a variety of out-of-school disadvantages and fall behind white children well before the first day of kindergarten (Gibbs and Downey 2020). If the score gap between black and white students grows in the years before school begins, it is hard to make sense of the same gap shrinking when school lets out for summer.

Gender gaps are another topic that has produced mixed summer findings. One recent study reported that girls gained math and reading skills faster than boys during summer, but not during school (Downey, Kuhfeld, and van Hek 2022). Yet an older study reported the opposite pattern for reading, with boys gaining more than girls during summer (Entwisle and Alexander 1994; Phillips and Chin 2004), and several other summer learning studies included gender, at least as a control variable, and found no sign of gender gaps in summer learning (Cooper et al. 1996; Downey et al. 2004; Entwisle and Alexander 1994; Kuhfeld et al. 2020; von Hippel and Hamrock 2019; von Hippel et al. 2018). In fact, authors of two summer learning manuscripts that initially included gender were instructed by reviewers to cut or minimize the gender findings because there was no significant seasonality to report (von Hippel and Hamrock 2019; von Hippel et al. 2018).

One recent summer study broadened its focus: instead of zooming in on gaps between specific groups of children, it zoomed out to look at total variance in learning rates (Condron, Downey, and Kuhfeld 2021). That study reported that the variance of learning rates was larger during summer than during school for reading, language, and science—but not for math. The inconsistency of the math finding was challenging to explain and challenging to reconcile with earlier studies that produced the opposite finding: in one earlier study, the variance of math scores grew during summers and shrank during school (von Hippel et al. 2018).

More broadly, it is not clear whether studies of seasonality in total variance can be expected to produce consistent results. Even on more basic questions, such as whether the total variance of test scores grows or shrinks with age, findings have been mixed for decades. On some tests, the variance of scores grows as children progress from kindergarten to eighth grade; on other tests, the variance holds steady or even shrinks (Clemans 1995; von Hippel and Hamrock 2019; Yen, Burket, and Fitzpatrick 1995a, b). If we cannot be sure whether total variance grows at all over a period of years, how can we be sure whether it grows faster over periods of a few months, during summer or during school?

Mixed results in the summer learning literature are just one example of the broader "replication crisis" or "generalizability crisis" in research. In fields from sociology and economics to medicine and psychology, a disquieting fraction of published findings cannot be replicated in new data, and some findings cannot even be reproduced in the data where they were originally reported (Christensen, Freese, and Miguel 2019; Ioannidis 2005; Open Science Collaboration 2015). In education, among interventions once thought to have "strong prior evidence of effectiveness," only 18 percent of replication attempts "found a statistically significant positive impact on [any] student academic outcome" (Boulay et al. 2018).

Although the nonreplicability of some summer learning results has begun to gain recognition, an impression may have taken hold that replicability problems are confined to older studies using outdated measurement methods (von Hippel 2019; von Hippel and Hamrock 2019). Some recent summer learning studies assert explicitly that, although older studies had limitations, studies using modern measurement tools should be more robust (Atteberry and McEachin 2021; Kuhfeld et al. 2020). More broadly, whenever studies of a single data set try to draw broad, general conclusions about patterns of learning and inequality, there is a tacit assumption that other data would produce similar results.

In this article, we show that the poor replicability of summer learning findings is neither a new problem nor, unfortunately, a problem that researchers have put in the rear-view window. In a historical review, we show that summer learning studies have produced mixed results for more than a century. Although some reasons for these mixed results are clear, others remain somewhat mysterious and hard to address. Next, in an empirical comparison of three recent data sets, we show that the problem of nonreplicable results is one that persists in modern data. In fact, hardly any recent claims about summer learning replicate across all three data sets. In the conclusion, we discuss implications and possible reasons for nonreplicable summer learning findings, settling on measurement issues as the most likely culprit. We make recommendations for future research, including reducing the field's reliance on proprietary or black-box tests that make it hard to tell what specific skills students retain or lose over the summer. We also call for a moratorium on making broad claims about the implications of summer learning for education and inequality without first verifying that the results supporting those claims replicate across more than one data set.

A Short History of Replication Failure in Summer Learning Research

Inconsistent results on summer learning go back more than 100 years. A meta-analysis by Cooper et al. (1996), often cited as though it simply found that summer learning loss exists, actually reported mixed or heterogeneous results throughout the history of summer learning research. Out of 80 early results obtained between 1906 and 1974, 48 results (60 percent) showed summer losses, but 26 (32 percent) showed summer gains, and six results (eight percent) showed no change over the summer (Cooper et al. 1996). Out of 52 later results, obtained between 1975 and 1994, 29 results (56 percent) found summer losses, but 21 results (40 percent) found summer gains, and two results (four percent) found no change over the summer (Cooper et al. 1996). Although a majority of studies showed summer losses, most studies were small and local, and the only large, nationally representative study that had been conducted at the time of Cooper et al.'s (1996) meta-analysis—the Sustaining Effects Study of nearly 120,000 first through sixth graders followed from the 1976-to-1977 school year to the 1978-to-1979 school year—reported that children lost little or no math skill, and actually gained reading skill, between spring and fall. When Cooper et al. (1996) averaged results across studies, summer learning was negative if the Sustaining Effects Study was excluded but zero or positive if the Sustaining Effects Study was included and given weight commensurate with its size.

Cooper et al.'s (1996) results were also mixed for summer learning gaps between children from advantaged and disadvantaged groups. On average, Cooper et al. found no summer learning gaps between boys and girls and no summer learning gaps between white and black children. Average summer reading losses were larger in studies of low-income children than in studies of middle-income children, but there were few studies that included both low- and middle-income children, and there was very little evidence available on income gaps in math. The Sustaining Effects Study, which tested both low- and middle-income children in both reading and math, did not fit the pattern of smaller studies; in the Sustaining Effects Study, income was not correlated with summer learning, or the correlation was no stronger during summer than during school (Ginsburg et al. 1981). Some later analysis and commentary debated this interpretation (Heyns 1987; Klibanoff and Haggart 1981), but the debate is difficult for a modern reader to adjudicate.[1]

Results published after Cooper et al.'s (1996) meta-analysis have agreed no better than results published before. Perhaps the most influential summer learning study of the past 40 years was the Beginning School Study (BSS), which started in the fall of 1982 with 838 first graders in 20 Baltimore public schools.[2] The BSS reported that reading and math gaps between low-SES and high-SES children grew only during summer, not during school, so that summer learning accounted for most of the SES gap by the end of elementary or middle school (Alexander, Entwisle, and Olson 2001)—a point repeated by the NSLA to this day (NSLA 2017). This finding, however, has not replicated particularly well in national data. The national Prospects Study, which followed children from first to second grade in 1992, found that most children gained reading and math skills between spring and fall. Although summer vocabulary gains were greater for high-SES children, summer gains in math concepts and applications were greater for low-SES children, and summer gains in reading comprehension and math concepts and applications were not significantly correlated with SES (Phillips and Chin 2004).

Results from two cohorts of the national Early Childhood Longitudinal Study—one that started with the kindergarten class of 1998 to 1999 (ECLS-K:1999) and one that started with the kindergarten class of 2010 to 2011 (ECLS-K:2011)—have also been mixed. Early results from the ECLS-K:1999 showed little or no loss of reading or math skills over the summer, on average. Early ECLS-K:1999 results did suggest that SES score gaps grew more quickly during summer than during school—but only in reading, not in math (Downey et al. 2004). As more ECLS-K:1999 data accumulated, the portrait of summer learning morphed—especially after test scores were rescaled in 2009. Across both ECLS-K cohorts (1999 and 2011), it now appears that SES gaps in reading and math grew little, if at all, during summer, or during school. Instead, gaps were present at the start of kindergarten and simply persisted through the end of elementary or middle school (von Hippel and Hamrock 2019; von Hippel et al. 2018).

Some recent summer learning studies have used data from NWEA (formerly the Northwest Evaluation Association), a commercial test vendor that tests children in fall and spring through the elementary grades (Atteberry and McEchin 2021; Downey et al. 2022; von Hippel and Hamrock 2019). NWEA data paint a portrait of summer learning loss that is not entirely compatible with evidence from other sources. Unlike the ECLS-K and ECLS-K:2011, NWEA data suggest that there are large average losses every summer (Atteberry and McEachin 2021; Downey et al. 2022; von Hippel and Hamrock 2019). Like the BSS, NWEA data suggest that SES gaps widen during the summer—but no faster than they widen during the school year (von Hippel and Hamrock 2019).

Methodological Reasons for Replication Failure

In short, over the past 100 years there has been little consistency in the results of summer learning studies. To some degree, mixed results may reflect real differences between the tested populations and what they learned or forgot during school and summer. But some of the discrepancies likely stem from design flaws and from differences in the measures and methods used in different studies.

One common flaw in summer learning studies is that students rarely take tests on the first and last day of summer. Instead, most students take spring tests weeks before school ends and fall tests weeks after school resumes. This flaw is just as common today as when it was first called out more than 40 years ago (Burkam et al. 2004; Cooper et al. 1996; Downey et al. 2004; Heyns 1987; Klibanoff and Haggart 1981; Phillips and Chin 2004). Nearly all older studies estimated summer learning by subtracting spring scores from fall scores, and these estimates were positively biased (Cooper et al. 1996). More recent studies, published since 2004, adjust summer learning estimates for school exposure (Burkam et al. 2004; Downey et al. 2004). These adjustments surely help, but no one knows how well they work. It would be better to test nearer the start and end of summer.

Another design flaw, especially common in older studies, is that spring and fall assessments tested different content (von Hippel 2019; von Hippel and Hamrock 2019). If a study's only goal were to estimate summer learning, it would test the same content in spring and fall, but instead test content often changes in the fall when students start a new grade. Despite efforts to equate scores across grades, it can be hard to distinguish changes in what tests cover from changes in what students know. This problem is especially acute when studies use fixed forms that ask all students the same questions. Adaptive tests, which ask different students different questions according to their abilities (Gershon 2005), may reduce the problem, but there will always be some tension between measuring summer learning and aligning a test with standards that change at the start of each grade.

Another measurement issue is test score scaling (von Hippel 2019; von Hippel and Hamrock 2019). Even when the content of a test is held constant, there are many ways to convert a pattern of right and wrong answers into a scaled test score, and different scaling methods can give very different impressions about skill growth and skill gaps. Gaps that appear to grow on one scale can appear to shrink on another one (Clemans 1993, 1995; Yen et al. 1995a, b). Over the history of summer learning research, learning has been estimated on a variety of scales including grade levels (Hayes and Grether 1969), the number or percentage of questions answered correctly (Downey et al. 2004), and Thurstone scales (Alexander et al. 2001). Modern studies increasingly use item response theory (IRT) scales, which in principle should make it easier to compare results across studies. In practice different IRT-scaled tests can produce very different-looking results (von Hippel and Hamrock 2019)—as we will see.

Measurement issues are not the only possible reason for discrepant estimates of summer learning. Summer learning estimates can be sensitive to model specification (Quinn 2015). Attrition and missing values can distort estimates of summer learning if students who test in spring do not test in fall, or vice versa.

Nor are test scores the only variable that can be measured in different ways. Many studies examine score gaps between high- and low-SES students, yet SES is a slippery construct, which different authors have measured using lunch-subsidy status, family income, parents' education, parents' occupational status, or some combination of these variables. Some studies measure the SES of individual families; others measure the SES of whole schools. Even when studies use the same variable—for example, family income—they may measure it in different ways or use different thresholds to define groups such as "low-income" families.

Finally, summer learning can be measured at different ages. Some studies are limited to the early grades; others run through the end of elementary or middle school. Although this could explain some of the differences between study results, the relationship between age and summer learning is another pattern that has not replicated well. Some studies report that summer losses increase with age (Cooper et al. 1996), but others report the opposite pattern (Alexander et al. 2001).

Our Contribution

For all these reasons, when summer learning studies reach different conclusions, it can be hard to know why. Study results may differ because of what models were fit, what tests were used, when tests were given, and how tests were scored—or because of what children were tested, how old they were, and how they were grouped.

In this study, we ask whether summer learning estimates become more replicable when we eliminate many common differences between studies. We compare three modern data sets, all of which tested children in the 2010s using adaptive tests scored with IRT scaling. We analyze each data set using the same statistical models and grouping children in the same way. We ask the following questions:

1. On average, how much reading and mathematics skill do children lose or gain during summer vacations?
2. Does the variance of test scores grow more quickly during summer vacations or during the school years?
3. On average, do the gaps between different student groups grow faster (or shrink slower) during the summer than during the school year?

We ask the last question about several different gaps: the gaps between girls and boys; the gaps between black, Hispanic, Asian American, and white children; and the gaps between children in high- and low-poverty schools. We focus on these gaps because they are the only ones that can be measured in all three data sets.

Despite eliminating many differences between studies, we find that hardly any findings in the summer learning literature replicate consistently across all three data sets. We discuss possible reasons for replication failure in the conclusion.

Data

We used data from three sources that measured children's skills in the early 2010s: the ECLS-K:2011 and data shared with us by two major test vendors, Renaissance Learning and NWEA.[3] Some previous summer learning studies have used the ECLS-K:2011 (Downey, Quinn, and Alcaraz 2019; Quinn and Le 2018; von Hippel et al. 2018), and some have used NWEA samples (Kuhfeld 2019; Kuhfeld et al. 2020; McEachin and Atteberry 2017; von Hippel and Hamrock 2019), although not the same NWEA sample as we use here. Renaissance Learning is a relatively new source for summer learning research.

All three sources included both fall and spring scores, permitting separate estimates of learning during school and during summer. All three sources avoided certain measurement artifacts that afflicted older studies. In particular, all three data sets used adaptive testing and scored assessments using IRT models that estimated each child's current skill level (often called *ability*[4] and represented by the Greek letter θ) while controlling for the difficulty of each test item. In all three sources, the resulting IRT scores were linear functions of the log odds that a child would give a correct response to an item of given difficulty (DeMars 2010). Items in the ECLS-K:2011 were developed to align with national standards and the standards of selected states; items on the Renaissance and NWEA items were developed to align with each state's standards, including a large pool of items that were common across states so that students from different states could be scored on the same scale. To facilitate comparisons across the three data sources, we standardized scores around the mean and standard deviation of the earliest grade available in each data set.

Table 1 gives descriptive statistics for each of the three sources. Specific details on each source are given below.

Federal Data: ECLS-K:2011

The ECLS-K:2011 began with a nationally representative sample of kindergartners attending U.S. public and private schools in the fall of 2010. It was a multistage cluster sample that sampled an average of 14 children within each school and sampled schools within primary sampling units, each of which was a large county or a group of similar and adjacent small counties. Asian Americans students were oversampled, and we used sampling weights to adjust descriptive statistics for oversampling and nonresponse, although we did not weight regression estimates because the weights were correlated with some regressors (Winship and Radbill 1994).

The ECLS-K:2011 contains a rich set of variables describing the children, families, teachers, and schools, but we used only the variables that were also available in our other data sources: the gender and race/ethnicity of each student and the percentage of students in each school who qualified for free or reduced-price lunch.

Tests: Content, Schedule, Adaptive Administration, and Scaling

Assessments designed for the ECLS-K:2011 covered content "derived from national and state standards, including those of the National Assessment of Educational Progress (NAEP), the ECLS-K:2011 frameworks, and selected state's curriculum standards" (National Center for Education Statistics 2021).

The ECLS-K:2011 assessed children's reading and math skills twice per year, in fall and spring, from kindergarten through second grade, then once per year, in spring only, from third through fifth grade. The schedule of tests was not optimal for a summer learning study. There were no fall tests after second grade, and even in grades with fall tests, students did not take tests near the first day of the school year, nor did they take spring tests near the last day of school year. Instead, students took the fall test seven weeks after school started, on average, and they took the spring test eight weeks, on average, before school ended. Students did not take tests at the same time. The total standard deviation of test dates was approximately three weeks in each fall and spring. Our estimates of school and summer learning adjust for variation in test dates and the differences between test dates and the first

Table 1: Characteristics of the three data sources

	Renaissance		NWEA		ECLS-K:2011
	Reading	Math	K–5 sample	2–5 sample	
Grades covered					
Lowest grade	1	1	K	2	K
Highest	5	5	5	5	2
Average dates					
First date of school	Aug. 22	Aug. 22	Aug. 22	Aug. 22	Aug. 22
Fall test date	Sept. 3	Sept. 9	Sept. 12	Sept. 18	Oct. 20
Weeks of school before fall test	2	3	3	4	8
Spring test date	May 9	May 13	April 27	May 2	April 18
Last date of school	June 4	June 4	June 4	June 4	June 4
Weeks of school after spring test	4	3	5	5	7
Student race/ethnicity					
White	50%	61%	74%	68%	53%
Black	23%	16%	2%	10%	13%
Hispanic	18%	14%	9%	12%	24%
Asian	5%	5%	4%	4%	4%
Other; multiethnic	5%	4%	10%	6%	6%
School characteristics					
Percentage of students qualifying for free or reduced-price lunch (school-level average)	61%	52%	48%	41%	45%
Sample size					
Test scores	127,160	90,606	32,262	122,295	22,310
Students	63,580	45,303	11,402	44,516	3,900
Schools	219	164	43	194	277

Notes: The first and last day of school were only recorded in the ECLS-K:2011; we assumed them to be the same, on average, in the NWEA and Renaissance data. ECLS-K:2011 statistics use the W3CF3P_30 sampling weight, and ECLS-K:2011 sample sizes were rounded to the nearest 10 in accordance with National Center for Education Statistics rules for restricted data. The number of test scores refers to math scores unless otherwise noted. In the ECLS-K:2011 data, the numbers of reading and math scores were nearly identical; in the NWEA data, the numbers of reading and math scores were exactly identical because the sample was limited to children with complete data for both.

and last days of schools, but these adjustments are not as good as having children actually take tests near the beginning and end of the school year.

Assessments were administered via a two-stage adaptive procedure. In the first stage, children took a short routing test. Results of the routing test determined which of three tests—easy, medium, or hard—the child would take in stage two (Tourangeau et al. 2015).

Assessments were scored using a three-parameter logistic IRT model, which estimated student skill net of the difficulty, guessability, and discrimination of each test item (Tourangeau et al. 2015).

Student skill was represented by a "theta" score estimating the log odds that a child would correctly answer an item of a given difficulty, guessability, and discrimination. Theta scores were more than 90 percent reliable. Because theta scores are on a log odds scale, they can take negative values, which many users find nonintuitive. To facilitate interpretation and comparison, we standardized the theta scores using the mean and standard deviation from the fall of kindergarten.

Sample Restrictions

The ECLS-K:2011 assessed all schools in the fall of kindergarten and the springs of kindergarten, first grade, and second grade but in the fall of first and second grade, assessments were limited to only a one-third random subsample of schools in first and second grade. We restricted our data to that subsample so that our estimates of summer learning would compare the same schools and children in spring and fall. We excluded tests taken in third, fourth, and fifth grade because those grades were tested only in spring and so could not separate school-year and summer learning. We also excluded the three percent of children who attended year-round schools, which had shorter summer vacations (von Hippel 2015).

After these restrictions, we had an analytic sample of 3,900 students who attended kindergarten in 280 schools. The first author (J.W.) accessed the ECLS-K:2011 using a restricted data license, which required us to round reported sample sizes to the nearest 10.

Vendor Data: Renaissance Learning and NWEA

We also analyzed data from Renaissance Learning and NWEA—two not-for-profit test vendors that assess students' skills in reading and math (as well as other subjects).[5] Renaissance Learning shared data with the second author (P.v.H), and NWEA shared data with the third author (J.M.). Data were restricted and not shared among authors, although code and results were.

In some ways, the vendor data were comparable to the ECLS-K:2011. Both vendor samples contained scores on reading and math tests given to large and diverse samples of children in the early 2010s—starting in fall 2011 in the NWEA data and fall 2013 in the Renaissance data.[6] Both vendors gave fall and spring tests throughout elementary school—not just in kindergarten through second grade, as in the ECLS-K:2011. Both vendors used adaptive testing methods and calculated IRT ability scores, although not in exactly the same way at the ECLS-K:2011.

But the similarity was limited. Although large and diverse, the vendor samples were not nationally representative like the ECLS-K:2011. And the vendors' demographic data were limited to the gender and race/ethnicity of individual students and the percentage of students at each school receiving free or reduced-price lunch. Note that, although both vendors had each school's percentage of students receiving free or reduced-price lunch, neither vendor had the lunch-subsidy status of individual students. This limited the demographic comparisons that we could make in the vendor data; to maintain comparability, we limited the ECLS-K:2011 to the same demographic variables. Note also that all our data sources measured lunch-subsidy status before the 2014 Community Eligibility Provision made free lunches universal at schools with more than 40 percent of students eligible (Hecht, Pollack Porter, and Turner 2020). Therefore, in our data the percentage of students receiving free or reduced-price lunch is a better measure of student-body poverty than it would be today.

A final difference was that the vendor tests had a different purpose than the tests given in the ECLS-K:2011. The ECLS-K:2011 tests were given purely for research purposes, but the vendor tests were purchased by schools, which used them to assess student progress. Note, however, that the NWEA and Renaissance tests were formative assessments that were often used voluntarily. They were not the high-stakes annual tests required of public schools in grades 3 through 8 under federal law.

Tests: Schedule, Adaptive Administration, Content, and Scaling

Vendor tests were aligned with state standards. Forty-one states use the Common Core State Standards, and states that rejected the Common Core often have standards that are not terribly different. Ninety-seven percent of NWEA assessment items are aligned with the Common Core (Set 2018), and using the same items in different states makes it easier to align scores across states.

The test schedule that schools used for vendor tests was not optimal for a summer learning study, but it was better than the test schedule in the ECLS-K:2011. Unlike the ECLS-K:2011, which tested only in the fall and spring of three grades (kindergarten through grade 2), NWEA tested in the fall and spring of six grades (kindergarten through grade 5), and Renaissance tested in the fall and spring of five grades (grades 1 through 5).[7] Students did not take vendor tests on the first and last day of the school year, but they tested closer to those dates than students in the ECLS-K:2011. In the fall, students took vendor tests two to four weeks after the school started, on average (vs. eight weeks in the ECLS-K:2011). In the spring, students took vendor tests three to five weeks before the end of school (vs. seven weeks in the ECLS-K:2011). The first and last day of school were only recorded in the ECLS-K:2011; we assumed them to be the same, on average, in the NWEA and Renaissance data. As in the ECLS-K:2011, students tested on different dates, and most of the variation in test dates lay between rather than within schools. Our estimates of school and summer learning adjust for variation in test dates and the differences between test dates and the first and last days of schools, but these adjustments are not as good as having children actually take tests at the beginning and end of the school year.

Like the ECLS-K:2011 tests, the vendor tests were adaptive, but the details were different. Whereas the ECLS-K:2011 tests adapted in two steps—with the result of a routing test determining the difficulty of the final test—the vendor tests were *continuously* adaptive, with each right or wrong answer affecting the difficulty of the next item.

Like the ECLS-K:2011 tests, the vendor tests estimated student skill using an IRT model, but again the details were different. The ECLS-K:2011 tests used a three-parameter IRT model that controlled for the difficulty, guessability, and psychometric discrimination of each item, whereas the vendor tests used a one-parameter IRT model (also known as a Rasch model) that only controlled for the difficulty of the item, assuming the guessability and discrimination of each item were the same (McCall, Kingsbury, and Olson 2004). To ensure that these assumptions were met, at least approximately, test vendors remove items that have unusually high or low discrimination or guessability. Vendors also try to remove items with cultural bias, both by conducting tests of differential item functioning and by getting input from a panel of experts representing different backgrounds.

Sample Restriction

Both vendors collect millions of scores every year, and some research has tried to use all of them, but attrition and selection make much vendor data unsuitable for seasonal longitudinal analysis. We imposed several restrictions on the data to ensure that the same schools and student bodies were being tested in the fall and spring of each grade.

First, although Renaissance and NWEA tests are sometimes given in preschool and middle school, they are most popular in elementary schools, so we limited the data to the elementary grades with the best coverage in reading and math. For the Renaissance data these were grades 1 through 5; for the NWEA data, they were kindergarten through grade 5, with a final test in the fall of grade 6.

Within these grades, not every student was tested. Some schools apparently used vendor tests only for selected students or selected grades. To ensure that the same student bodies were represented in every grade, we limited the data to schools that tested approximately the same number of students in every grade—grades 1 through 5 in the Renaissance data, or kindergarten through grade 5 in the NWEA data.[8] In the NWEA data, limiting to schools that tested every grade from kindergarten through grade 5 ruled out many schools with substantial African American enrollments, so we also used a less-limited NWEA data set that only required schools to test equal numbers of students in grades 2 through 5. Our tables show results for both the kindergarten through grade 5 and the grades 2 through grade 5 NWEA samples, side by side. Our main figures show results for the kindergarten through grade 5 NWEA sample, but figures for the grades 2 through 5 NWEA sample are reported in the online supplement.

In contrast to the ECLS-K:2011, in vendor data it is challenging to follow students for multiple years. Among students who took NWEA tests in the fall of 2018, one-quarter did not take NWEA tests one year later (Johnson and Kuhfeld 2020). Likewise, in our Renaissance data, among first graders tested in the 2013-to-2014 school year, fewer than half were still taking Renaissance tests as sixth graders in the 2018-to-2019 school year. There are several reasons for this. Renaissance and NWEA tests are not high-stakes tests required by state governments but low-stakes formative assessments that schools and distracts use at their discretion. Every year some districts let their contracts with NWEA or Renaissance expire, whereas other schools start using vendor tests for the first time. Some schools use vendor tests only for selected students or in selected grades, and students can move from a school that uses Renaissance or NWEA tests to a school that does not.

To reduce the problem of attrition, instead of following students longitudinally across multiple school years, we used an "accelerated" longitudinal design that simultaneously followed six cohorts of students for one year each (Galbraith, Bowden, and Mander 2017). Each cohort was tested in the fall and spring of one school year (2011 to 2012 in the NWEA data) and then the fall of the next school year (2012 to 2013). One cohort was followed from kindergarten into first grade, one was followed from first grade into second grade, and so on.

Finally, although NWEA collected student gender and race/ethnicity for almost every child, Renaissance did not require these variables, and some schools did not report them. We dropped from the Renaissance data all schools not reporting student gender and race/ethnicity.

Table 1 summarizes the sample size, demographic characteristics, and test schedules of our three data sources. Notice that, even after our restrictions, the vendor data had more test scores and children (although not more schools) than our ECLS-K:2011 sample. Notice also that all three samples are considerably larger and more diverse than the Beginning School Study conducted of 838 children at 20 Baltimore public schools (Alexander et al. 2001), which since the 1990s has exerted outsized influence on our thinking about summer learning patterns.

Methods

We graphed average test scores against average test dates in the fall and spring of each school year. Although these graphs conveyed important insights, they gave a biased estimate of summer learning because the tests were not given on the first and last day of summer. To reduce bias, we linearly extrapolated beyond the test dates to the scores that would have been obtained on the first and last day of summer if children continued to learn at the same rate before and after the tests. Our growth models also extrapolated beyond the test dates, as we will explain next.

Growth Models

We specified two models: (1) a model of *average* scores and growth rates and (2) a model of *gaps* between the average scores and growth rates of different groups. To keep things simple, we will describe models for kindergarten through grade 2. It is straightforward to extend the models to a longer grade range.

Average Model

Our average model was

$$Y_{sct} = \alpha_{0sc} + \beta_0 G0_{sct} + \alpha_1 S1_{sct} + \beta_1 G1_{sct} + \alpha_2 S2_{sct} + \beta_2 G2_{sct} + e_{sct}.$$

Here Y_{sct} was the standardized reading or math score of child c in school s on test occasion $t = 1, \ldots, 6$. At the time of the test, $G0_{sct}, G1_{sct}, G2_{sct}$ were the months that the child had spent in grades 0 (kindergarten), 1, and 2, and $S1_{sct}, S2_{sct}$ were the number of months that the child had been spent in summer vacations before first and second grade (summer 1 and summer 2). Because different children were tested on different dates, months in school and summer depended on the test date. For example, if a first grader's school opened on August 22, 2011, and gave a fall test on October 20, 2011, then at the time of the fall test the child had $G0_{sct} = 9.3$ months in kindergarten, $S1_{sct} = 2.7$ months in summer 1, $G1_{sct} = 1.0$ months in first grade, and no time yet in summer 2 or second grade ($S2_{sct} = G2_{sct} = 0$).

Therefore, the parameters $\beta_0, \beta_1, \beta_2$ were average monthly learning rates during kindergarten, first, and second grade, and the parameters α_1, α_2 were average monthly learning rates during summers 1 and 2. All learning rates were measured in standard deviations per month, where the standard deviation is that of the first test given.

The model included a random child-level intercept α_{0sc} to account for the correlation among tests given to the same child, and school-clustered standard errors to account for correlations among children in the same school. In the accelerated cohort design, which we used for the vendor data, the model also included a dummy variable for each cohort.

The model was linear, meaning that it assumed that children learned at a constant rate throughout the school year, so that the scores that would have been achieved on the first and last day of school could be obtained by linearly extrapolating beyond the scores that were actually achieved on the test dates. Some of our estimates may be sensitive to extrapolation, especially when the difference between the test date and the first or last day of school was large. Although the linear model was only an approximation, there was no way to evaluate nonlinearity with only two test scores per student per year. As discussed earlier, this problem is very common because none of the major summer learning studies tested children near the first and last day of school (Downey et al. 2004; Entwisle and Alexander 1992; Lee and Burkam 2004).

We combined model parameters to compare learning rates across school years and summers. The average school-year learning rate was $\widehat{\beta} = (\beta_0 + \beta_1 + \beta_2)/3$, and the average summer rate was $\widehat{\alpha} = (\alpha_1 + \alpha_2)/2$. Then $\widehat{\alpha}/\widehat{\beta}$ was the ratio of school to summer learning rates. With the average summer vacation lasting 2.65 months, 2.65 $\widehat{\alpha}$ was the average amount gained or lost over summer vacation, and the ratio 2.65 $\widehat{\alpha}/\widehat{\beta}$ represented summer gains or losses expressed as the average months of school-year learning gained or lost over summer vacation. Standard errors of estimated averages and ratios were estimated using the delta method.

Gap Model

Our model of gaps was similar except that time in school and summer could interact with an X variable representing some child- or school-level characteristic:

$$\begin{aligned} Y_{sct} = \alpha_{0cs} &+ \beta_0 G0_{sct} + \alpha_1 S1_{sct} + \beta_1 G1_{sct} + \alpha_2 S2_{sct} + \beta_2 G2_{sct} \\ &+ \gamma_0 X_{sc} + \delta_0 G0_{sct} X_{sc} + \gamma_1 S1_{sct} X_{sc} + \delta_1 G1_{sct} X_{sc} \\ &+ \gamma_2 S2_{sct} X_{sc} + \delta_2 G2_{sct} X_{sc} + e_{sct}. \end{aligned}$$

To interpret the coefficients concretely, suppose X_{sc} was a dummy for female gender. Then the coefficient δ_0 was the score gap between girls and boys on the first day of kindergarten; the parameters $\delta_0, \delta_1, \delta_2$ were the monthly rate by which the gap grew or shrank during kindergarten, first, and second grade; and the parameters γ_1, γ_2 were monthly gap changes during summers 1 and 2. We combined these parameters to compare gap changes during school and summer. In particular, $\widehat{\delta}/\widehat{\gamma}$ was the ratio of monthly gap changes during school and summer, where $\widehat{\delta} = (\delta_0 + \delta_1 + \delta_2)/3$ is the average monthly gap change during school and $\widehat{\gamma} = (\gamma_1 + \gamma_2)/2$ is the average monthly gap change during summer.

More Complex Models, and Why They Are Unnecessary

Although not simple in any absolute sense, our models are simpler than the models fit in some past summer learning studies. Our basic conclusion turns out to be that even the simplest results fail to replicate. Simple models, and even simple graphs, suffice to show this. Fitting more complicated models would not change the conclusion and might obscure it.

One of our simplifying decisions was to include only one child or school variable at a time. Some past studies have done this (Quinn 2015; von Hippel and Hamrock 2019), but others have included all child and school variables at once (e.g., Downey et al. 2004; von Hippel et al. 2018). Models with more variables have more complicated interpretations. For example, a model that includes only race estimates the average gap between black and white students, whereas a model that included race and school poverty would estimate the average gap between black and white students attending schools with similar poverty levels. Because more complicated models are not more replicable, we chose to keep our models simple.

Another simplifying decision was to focus on average learning rates, learning gaps, and the standard deviation of test scores. The model parameters $\alpha, \beta, \gamma, \delta$ represent average learning rates and learning gaps between different *groups*, but the learning rates of *individual* students may deviate substantially from these averages. Some previous studies tried to model individual differences by adding school- and child-level random effect to the parameters representing learning rates (Downey et al. 2004; von Hippel and Hamrock 2019; von Hippel et al. 2018). Unfortunately, such a random slopes model could not be fit to an accelerated cohort design like the one that we used in the NWEA and Renaissance data.[9]

Results

This section examines how well summer learning patterns replicate across our three data sets.

Average Summer Losses

How much did students lose over the summer, on average? Figure 1 graphs average scores on the dates of the fall and spring reading and math tests (black lines) and then extrapolates (blue lines) to the scores that would have been obtained on the first and last day of school if children kept learning at a similar rate before and after the tests.

Whether we extrapolated or not, the data sources disagreed about how much, if anything, children's skills declined during the summers. In reading, summer losses appeared dramatic on the NWEA tests, but summer losses appeared milder on the Renaissance tests and practically nonexistent on the ECLS-K:2011 tests. In

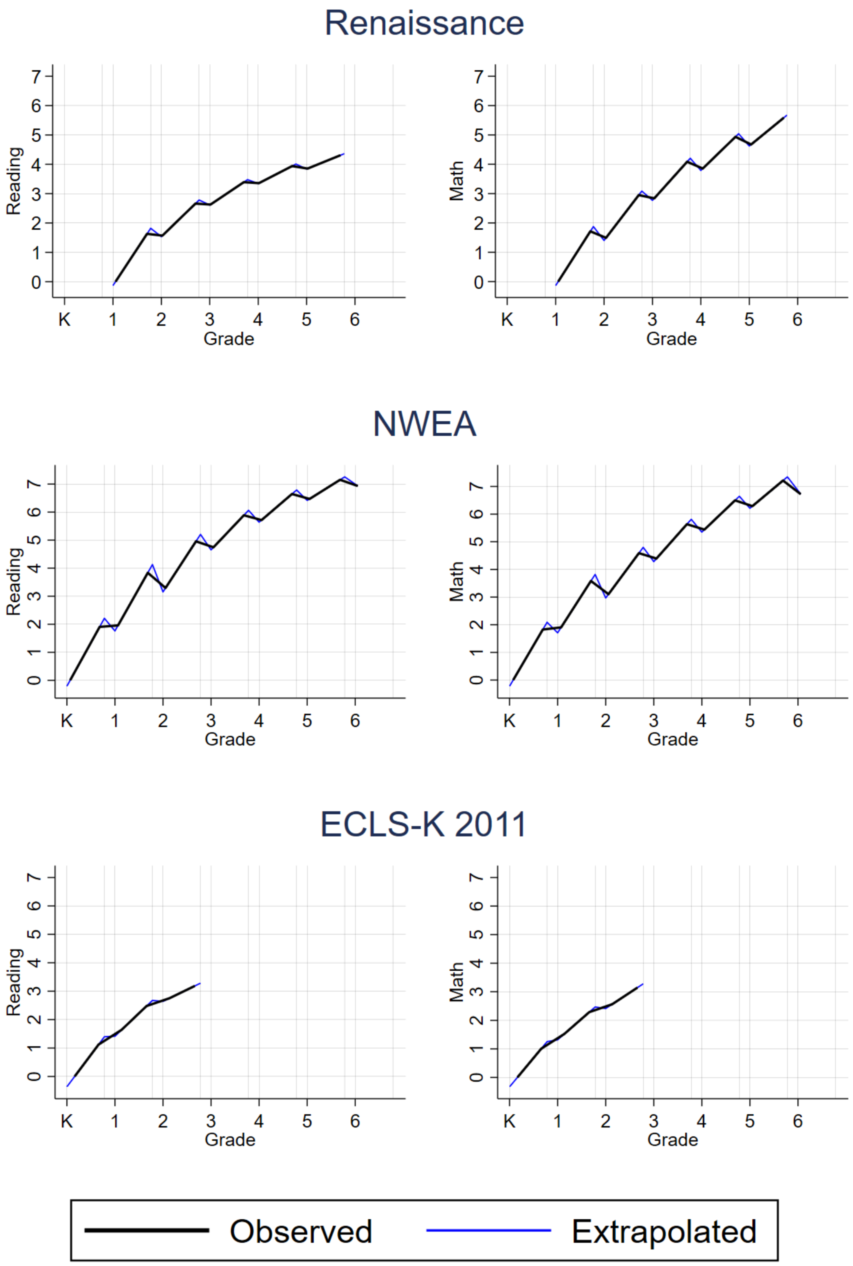

Figure 1: Average school and summer gains and losses in reading and math in three different data sets. The black lines connect the average of observed scores in fall and spring. The blue lines connect the scores that, according to linear extrapolation, would have obtained on the first and last day of each grade.

math, summer losses appeared substantial on the NWEA and Renaissance tests but practically nonexistent on the ECLS-K:2011 tests.

There was no agreement on which subject was most prone to summer loss. The Renaissance data showed larger summer losses in math than in reading, but the NWEA data showed similar summer losses in math and reading, and the ECLS-K:2011 showed practically no summer losses in either subject.

These disagreements are hard to reconcile. Although each data source covers a different range of summers, disagreements persist even when we compare the same summers. For example, the NWEA tests showed their largest losses in the summer after first grade—a summer when the Renaissance tests showed only small losses, and the ECLS-K:2011 showed no losses at all.

Extrapolating beyond the test dates did not account for the discrepancies, either. Whether we extrapolated or not, summer losses appeared much larger on the NWEA tests than on the ECLS-K:2011 tests. Whether we extrapolated or not, summer losses on the NWEA and Renaissance tests appeared similar in math but less similar in reading.

Table 2 confirms these discrepancies with estimates of monthly gains and losses from a linear growth model. On the NWEA tests, children lost between 0.3 and 0.5 standard deviations (SD) over the summers, an amount equivalent to two to three months of school-year learning; in other words, children lost skills as quickly during the summer as they gained skills during the school year. On the ECLS-K:2011 tests, however, summer losses were trivial—statistically insignificant in math and just 0.03 SD in reading, an amount equivalent to less than a week of school-year learning. On the Renaissance tests, summer losses in math were as large as on the NWEA tests, but summer losses in reading were smaller.

Changes in the Standard Deviation during School and during Summer

What about the SD of test scores? Did the SD grow more quickly during summer vacations than during the school years? Table 3 answers this by giving the SD in the fall and spring of each school year.

Across tests, there was no agreement about when the SD grew fastest, or if it grew at all. In math, for example, the SD of the NWEA test grew during school and shrank during summer, but the SD of the ECLS-K:2011 test *shrank* during school and *grew* during summer, whereas the SD of the Renaissance test grew slightly during both school and summer. Considering the three tests together, there was no consistency about whether the SD of math scores grew faster during school or summer and therefore no consistent message about whether inequality came primarily from inside or outside schools.

In reading, the three tests also disagreed about whether and when the SD grew, although the details of the disagreement were different.

Although different tests covered different grades, that did not explain the disagreement. Even if we focused on just kindergarten and the summer after—a period covered by both the NWEA and the ECLS-K:2011—results for the SD still disagreed. In math, the SD of the NWEA test grew by 20 percent during kindergarten and

Table 2: Average summer and school-year learning. Estimates from linear growth models

	Reading				Math			
	Renaissance	NWEA (K–5)	NWEA (2–5)	ECLS-K:2011	Renaissance	NWEA (K–5)	NWEA (2005)	ECLS-K:2011
Monthly SDs gained during school year	$0.114^\dagger$	$0.178^\dagger$	$0.136^\dagger$	$0.132^\dagger$	$0.161^\dagger$	$0.182^\dagger$	$0.150^\dagger$	$0.129^\dagger$
	(0.001)	(0.005)	(0.002)	(0.001)	(0.002)	(0.004)	(0.002)	(0.001)
Monthly SDs lost during summer	$-0.078^\dagger$	$-0.189^\dagger$	$-0.110^\dagger$	-0.011^*	$-0.161^\dagger$	$-0.187^\dagger$	$-0.146^\dagger$	−0.003
	(0.005)	(0.022)	(0.007)	(0.004)	(0.006)	(0.016)	(0.005)	(0.004)
Total SDs lost during summer	$-0.207^\dagger$	$-0.498^\dagger$	$-0.290^\dagger$	-0.029^*	$-0.424^\dagger$	$-0.493^\dagger$	$-0.386^\dagger$	−0.009
	(0.012)	(0.058)	(0.018)	(0.012)	(0.015)	(0.043)	(0.014)	(0.011)
Ratio of monthly summer loss to monthly school-year gain	$-0.688^\dagger$	$-1.060^\dagger$	$-0.812^\dagger$	-0.083^*	$-0.999^\dagger$	$-1.027^\dagger$	$-0.976^\dagger$	−0.025
	(0.034)	(0.100)	(0.041)	(0.033)	(0.028)	(0.069)	(0.028)	(0.033)
Months of school-year learning lost during summer	$-1.814^\dagger$	$-2.796^\dagger$	$-2.141^\dagger$	-0.219^*	$-2.633^\dagger$	$-2.709^\dagger$	$-2.573^\dagger$	−0.066
	(0.089)	(0.264)	(0.109)	(0.088)	(0.074)	(0.181)	(0.073)	(0.087)

Notes: Standard errors in parentheses. † $p < 0.01$; $* \ p < 0.05$.

Table 3: SDs of test scores in fall and spring

		Reading				Math			
Grade	Season	Renaissance	NWEA (K–5)	NWEA (2–5)	ECLS-K:2011	Renaissance	NWEA (K–5)	NWEA (2–5)	ECLS-K:2011
K	Fall		1.00		1.00		1.00		1.00
	Spring		1.31		0.91		1.20		0.80
1	Fall	1.00	1.31		0.91	1.00	1.16		0.88
	Spring	1.03	1.42		0.87	0.99	1.25		0.89
2	Fall	1.01	1.64	1.63	0.78	1.01	1.18	1.16	0.89
	Spring	0.91	1.38	1.47	0.75	1.02	1.09	1.11	0.84
3	Fall	0.95	1.56	1.57		1.04	1.13	1.12	
	Spring	0.89	1.40	1.43		1.03	1.11	1.14	
4	Fall	0.91	1.51	1.50		1.06	1.17	1.16	
	Spring	0.90	1.34	1.37		1.14	1.25	1.23	
5	Fall	0.91	1.44	1.44		1.14	1.24	1.26	
	Spring	0.91	1.36	1.39		1.22	1.39	1.40	
6	Fall		1.49	1.46			1.30	1.30	
Average % change in SD									
Summers		1%	9%	5%	–6%	2%	–1%	–1%	5%
School years		–3%	0%	–8%	–5%	3%	6%	4%	-8%
% change in SD per month									
Summers		1%	3%	2%	–2%	1%	0%	0%	2%
School years		–0.3%	0%	–1%	–1%	0.3%	1%	0%	–1%

shrank by three percent during summer, but the SD of the ELCS-K:2011 test *shrank* by 20 percent during kindergarten and *grew* by 10 percent during summer. In reading, the SD of the NWEA test grew by 30 percent during kindergarten and held steady during summer, but the SD of the ELCS-K:2011 reading test shrank by nine percent during kindergarten and held steady during summer.

Even if we focused on a single test vendor, results still disagreed across subjects. For example, on the NWEA tests, the reading SD grew during summer and shrank or held steady during school, but the math SD *shrank* (just slightly) during summer and *grew* during school. Looking at the NWEA results alone, it would be tempting to conclude that reading inequality came primary from home and math inequality came primarily from school. But that pattern did not replicate on the Renaissance and ECLS-K:2011 tests.

The tests did not just disagree on whether the SD grew faster during school or summer. They also disagreed on a more basic question: did the SD grow over time at all? Or did it shrink? In reading, for example, between the fall of kindergarten and the spring of second grade, the SD grew by 38 percent on the NWEA test but shrank by 25 percent on the ECLS-K:2011 test. In math, over the same period, the SD grew by nine percent on the NWEA test but shrank by 16 percent on the ECLS-K:2011 test. Although the disagreements were strongest between the NWEA and the ECLS-K:2011 tests, there were also disagreements between the NWEA tests and the Renaissance tests. For example, in reading, between the fall of first grade and the spring of fifth, the Renaissance SD shrank by nine percent, but the NWEA SD grew by four percent. Because the tests disagreed on whether the SD grew at all, the question of whether the SD grows faster during school or summer seems moot.

Gaps between High-Poverty and Low-Poverty Schools

Figure 2 graphs average trends in reading and math score gaps between high-poverty and low-poverty schools.

On balance, the gaps between high-poverty and low-poverty schools changed very little as children grew older. On the NWEA and ECLS-K:2011 tests, the gaps were no larger (in fact a little smaller) at the end of second grade than they were at the start of kindergarten. On the NWEA and Renaissance tests, the gaps were about the same at the end of sixth grade as at the start of first. There were differences in some particulars—for example, the Renaissance test showed slight growth in the math gap, but the NWEA did not—but for the most part the results showed little net change in gaps during elementary school.

To the degree that the gaps changed, the three data sources disagreed about the pattern of changes across school and summer. On the NWEA tests, there was a regular, sawtooth, seasonal pattern, with gaps shrinking in every school year and growing in every summer except one. On the other tests, however, any changes in the gap display less regular patterns that rarely agree with the NWEA. On the Renaissance reading test, for example, the gap *grew* during first grade and *shrank* during the next summer—the opposite of the NWEA pattern—and after that there was little seasonality, with the gap persisting more or less unchanged through the end of fifth grade. Although there were occasional summers when the gap grew

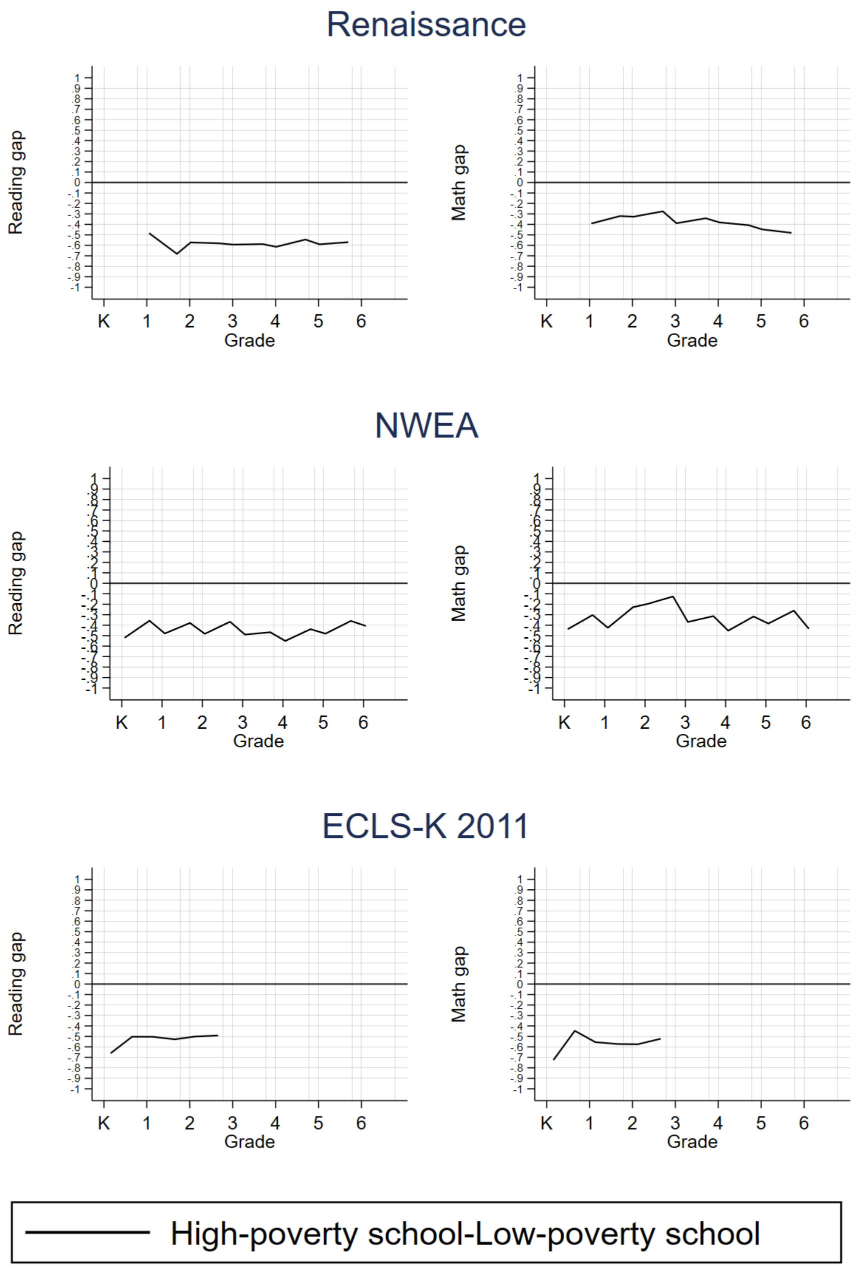

Figure 2: Average gap between high-poverty and low-poverty schools (i.e., more than 40 percent vs. less than 40 percent free and reduced-price lunch).

and occasional school years when the gap shrank, any seasonal patterns were not consistent across the years or across the tests.

It seems very difficult to draw general conclusions about when or whether the gaps between high-poverty and low-poverty schools grow or shrink. Almost any conclusion would be nonreplicable. The one exception may be during kindergarten, when both the NWEA and the ECLS-K:2011 show gaps shrinking in both reading and math, although kindergarten scores were not available from Renaissance.

Table 4 uses our linear growth model to check the conclusions that we drew by inspecting the graphs. In reading, our conclusions are confirmed. Only the NWEA data show any significant difference between school and summer gap change. In fact, only the NWEA data show any significant change in the gap at all. On the NWEA reading test, in the large analytic sample that included grades 2 through 5, the gap between high-poverty and low-poverty schools grew by approximately 0.04 SD per month during summer and shrank by approximately 0.01 SD per month during school. Both these estimates were statistically significant, and so was the difference between them (all $p < 0.05$). Estimates in the smaller analytic sample, which included kindergarten through grade 5, were similar in direction but not statistically significant. Neither the Renaissance data nor the ECLS-K:2011 data showed any significant difference between school and summer gap growth; in fact, neither showed any significant gap growth during summer, although the ECLS-K:2011 did show modest gap shrinkage, averaging 0.01 SD per month, during the school year.

In math, the results were somewhat different. In math, it was the ECLS-K:2011 that showed significant gap growth during summer (approximately 0.03 SD per month), significant gap shrinkage during school (0.16 SD per month), and a significant difference between school and summer gap changes (all $p < 0.01$). Results for the kindergarten through grade 5 NWEA sample were directionally similar, but nonsignificant; results for the larger grades 2 through 5 NWEA sample showed only trivial and nonsignificant changes in the gap during school or summer (all estimates less than 0.01 SD per month, no $p < 0.05$). The Renaissance data showed significant gap growth during summer (0.02 SD per month, $p < 0.05$) but no significant difference between school and summer gap growth.

Gaps between Girls and Boys

Figure 3 graphs trends in the average score gaps between girls and boys. Gender gaps were smaller than other gaps examined in this study. On reading tests, girls scored between 0.1 and 0.3 SD ahead of boys, depending on the test and age. On math tests, girls scored between 0.1 ahead of boys and 0.2 SD behind boys.

For the most part, gender gaps showed little net change as children grew older. In math, the NWEA and ECLS-K:2011 tests showed girls starting at parity or slightly ahead, then falling about 0.1 SD behind as they grew older. On the Renaissance math tests, however, girls and boys started at parity, and girls pulled nearly 0.1 SD *ahead* as they grew older.

In reading, the NWEA and Renaissance tests showed girls' lead growing and then shrinking, so that girls were no further ahead at the end of fifth or sixth grade

Table 4: Summer and school-year change in score gaps

	Reading				Math			
	Renaissance	NWEA (K–5)	NWEA (2–5)	ECLS-K: 2011	Renaissance	NWEA (K–5)	NWEA (2–5)	ECLS-K: 2011
Gap between high- and low-poverty schools								
Monthly gap change, summer	−0.002	−0.051	−0.038†	−0.002	−0.021*	−0.053	−0.009	−0.032†
	(0.009)	(0.037)	(0.013)	(0.009)	(0.011)	(0.028)	(0.011)	(0.009)
Monthly gap change, school	−0.002	0.013	0.009*	0.009†	−0.001	0.014	−0.004	0.016†
	(0.002)	(0.009)	(0.004)	(0.003)	(0.003)	(0.007)	(0.004)	(0.003)
Difference in monthly gap change,	−0.001	−0.064	−0.047†	−0.011	−0.020	−0.067	−0.004	−0.048†
summer minus school	(0.011)	(0.046)	(0.017)	(0.012)	(0.013)	(0.035)	(0.014)	(0.011)
Gap between girls and boys								
Monthly gap change, summer	0.004	0.020*	0.027†	0.001	0.027†	0.021†	0.017†	−0.000
	(0.004)	(0.008)	(0.005)	(0.006)	(0.005)	(0.007)	(0.003)	(0.005)
Monthly gap change, school	−0.000	−0.004	−0.010†	0.002	−0.005†	−0.010†	−0.005†	−0.004*
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Difference in monthly gap change,	0.005	0.024*	0.038†	−0.002	0.032†	0.030†	0.023†	0.003
summer minus school	(0.004)	(0.010)	(0.006)	(0.007)	(0.005)	(0.008)	(0.003)	(0.007)
Gap between Asian and white children								
Monthly gap change, summer	−0.003	0.149†	0.077†	0.022*	0.013	0.107†	0.040†	−0.004
	(0.010)	(0.022)	(0.011)	(0.011)	(0.011)	(0.039)	(0.010)	(0.012)
Monthly gap change, school	−0.005†	−0.027†	−0.016†	−0.006	0.000	−0.017	0.006*	0.002
	(0.002)	(0.007)	(0.003)	(0.004)	(0.003)	(0.009)	(0.003)	(0.003)
Difference in monthly gap change,	0.002	0.175†	0.094†	0.028*	0.012	0.124†	0.034†	−0.007
summer minus school	(0.011)	(0.029)	(0.014)	(0.014)	(0.013)	(0.048)	(0.012)	(0.015)
Gap between black and white children								
Monthly gap change, summer	−0.003	0.033	0.015	0.024*	0.028†	0.042	0.025†	−0.001
	(0.009)	(0.027)	(0.010)	(0.011)	(0.010)	(0.024)	(0.009)	(0.010)
Monthly gap change, school	−0.006†	−0.011	−0.006*	−0.005	−0.016†	−0.015*	−0.011†	−0.005
	(0.002)	(0.006)	(0.003)	(0.003)	(0.003)	(0.008)	(0.003)	(0.003)
Difference in monthly gap change,	0.003	0.044	0.020	0.029*	0.044†	0.058	0.036†	0.004
summer minus school	(0.012)	(0.032)	(0.012)	(0.014)	(0.012)	(0.031)	(0.012)	(0.012)
Gap between Hispanic and white children								
Monthly gap change, summer	0.009	0.048	−0.027*	0.016	−0.049†	0.080†	0.010	−0.034†
	(0.008)	(0.027)	(0.014)	(0.009)	(0.012)	(0.022)	(0.009)	(0.008)
Monthly gap change, school	−0.006†	−0.018*	0.014†	0.004	0.008†	−0.021†	−0.005*	0.017†
	(0.002)	(0.008)	(0.004)	(0.003)	(0.003)	(0.005)	(0.002)	(0.002)
Difference in monthly gap change,	0.015	0.066	−0.041*	0.011	−0.057†	0.102†	0.015	−0.051†
summer minus school	(0.009)	(0.033)	(0.017)	(0.011)	(0.014)	(0.026)	(0.011)	(0.010)

Notes: In rows summarizing gap changes between group A and B, positive values mean that group A learned faster than group B, and negative values mean that group B learned faster than group A. In rows summarizing the difference between summer gap change and school gap change, negative values mean that the gap grew faster during summer than during school (if it grew during school at all). Standard errors in parentheses. † $p < 0.01$; ∗ $p < 0.05$.

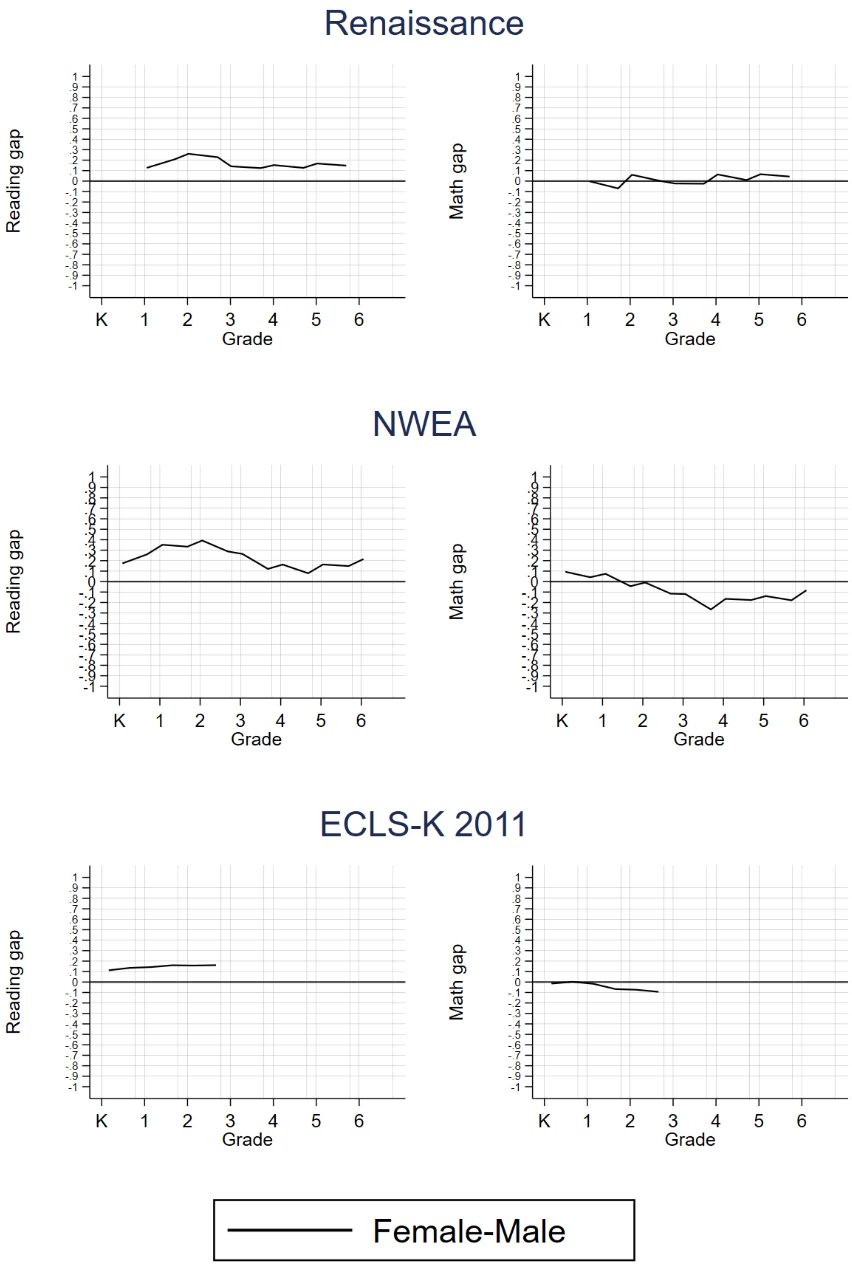

Figure 3: Average gaps between boys and girls.

than they were at the start of first grade or kindergarten. In the ECLS-K:2011 girls' lead grew just slightly between the start of kindergarten and the end of second grade.

To the extent that gender gaps changed, there was partial but incomplete agreement about how gaps changed across school and summer. In both reading and math, the NWEA and Renaissance tests showed girls gaining on or pulling ahead of boys during most but not all summers, and boys regaining lost ground during most but not all school years. On the ECLS-K:2011 tests, however, no seasonality in gender gaps was evident.

Table 4 summarizes these patterns using contrasts from a linear growth model. In reading, the NWEA test shows that girls gained 0.02 to 0.03 SD per month more than boys during the summer, but girls lost up to 0.01 SD per month more than boys during the school year. Both these estimates are statistically significant, and so is the difference between them (all $p < 0.01$). However, neither the Renaissance nor the ECLS-K:2011 reading tests show any significant difference between school and summer changes in the gender gap.

In math, all three data sets showed that boys learned significantly more than girls during the school year, but boys' advantage was small, ranging from 0.004 to 0.010 SD per month. But the data sets did not agree on gender patterns during summer; the Renaissance and NWEA tests showed that girls gained 0.02 to 0.03 SD per month more than boys during summer, but the ECLS-K:2011 math test showed no summer change in the gender gap.

Gaps between Children of Different Races and Ethnicities

Next, we compare the fall and spring scores of children who are Asian American, Hispanic, black non-Hispanic, or white non-Hispanic. Figure 4 graphs trends in the average score gaps between non-Hispanic whites and every other group. On nearly every test occasion, Asian American children were ahead of non-Hispanic white children, and non-Hispanic white children were ahead of children who are African American or Hispanic.

On the NWEA tests, the gap between Asian American and white children appears highly seasonal, especially in reading. The gap grew dramatically during the summers and shrank just as dramatically during the school years. But on the other tests, the gap between Asian American did not grow and shrink seasonally, and on none of the tests, including the NWEA, did the gaps between white and black or Hispanic children display any seasonality.

Table 4 summarizes these patterns using contrasts from a linear growth model. In reading, both the NWEA and ECLS-K:2011 tests showed Asian Americans gaining more than white students during summer and less than white students during school. Both these estimates were statistically significant, and so was the difference between them (all $p < 0.01$). However, the Renaissance tests did not show the same pattern. In math, only the NWEA tests showed Asian Americans gaining more than white students during summer and less during school; neither the ECLS-K:2011 nor the Renaissance tests show that pattern.

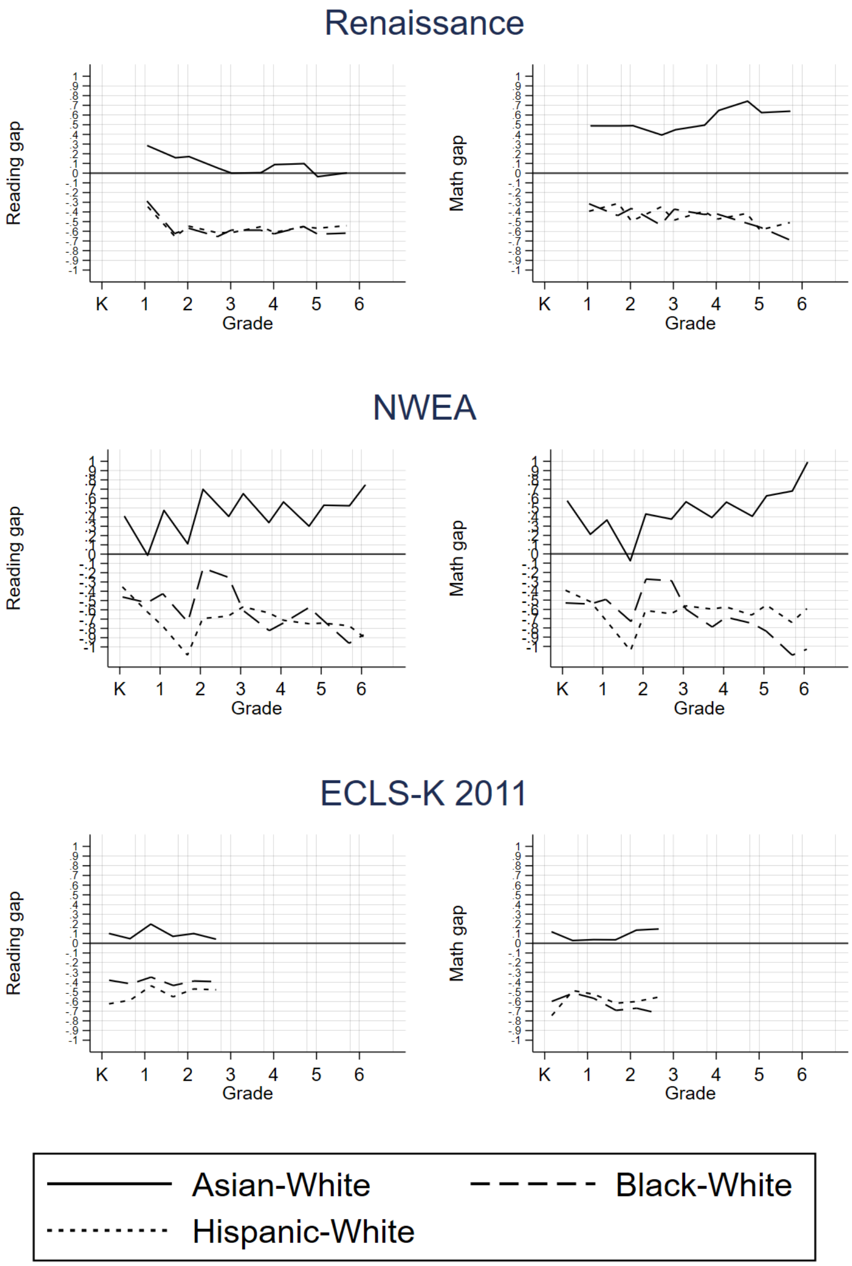

Figure 4: Average gaps between non-Hispanic white children and children of three other races/ethnicities.

Half the results comparing black and white children suggested that black children learned significantly faster than white children during the summer, but significantly slower during the school year. This pattern occurred more than once, but it did not replicate consistently. In math, the pattern was evident on the Renaissance and NWEA tests, but not in the ECLS-K:2011 test. In reading, the pattern was evident on the ECLS-K:2011 test, but not on the Renaissance and NWEA tests.

Taken at face value, the finding that black children learn a little slower during the school year is not hard to reconcile with other findings. For example, black children tend to have less experienced teachers (Goldhaber, Lavery, and Theobald 2015) and are less likely to have the benefit of a teacher of their own race (Redding 2019). However, the finding that black children learn faster during the summer is harder to explain, although it is not unprecedented (von Hippel and Hamrock 2019; von Hippel et al. 2018). Black families often lack the financial and educational resources needed to support children's learning when school is out, and this is likely why black children score lower than white children from the first day of kindergarten (Redding 2019). Why, then, would black children catch up during summer? Although the finding is puzzling, there may be risk in overinterpreting it, because it fails to replicate half the time.

For the Hispanic–white gap, the seasonal pattern is even less consistent. In math, the Renaissance and ECLS-K:2011 tests show Hispanic children learning significantly less than white children during summer and significantly more than white children during school—but one NWEA sample show the opposite pattern. In reading, one NWEA sample shows Hispanic children learning significantly more than white children during school and less than white children during summer—but this pattern does not replicate on the Renaissance or ECLS-K:2011 tests.

Discussion

Few Results Replicated

We found only two simple patterns that replicated consistently across data sources:

1. Substantial score gaps were already present when children began elementary school. Evidently a large share of cognitive inequality originated before school began.
2. During summer, most children made little or no progress, but during the school year they learned relatively quickly. Evidently schools accelerate learning, as they should.

Beyond that, there were practically no summer learning results that replicated.

There was little agreement even on the simplest question: how much, if anything, did students lose over the summer? On NWEA tests, summer loss appeared disastrous: children appeared to lose skills during summer as quickly as they gained skills during the school year. But on the ECLS-K:2011 tests it appeared that children lost little or nothing, on average, over the summer. On Renaissance tests, summer math losses looked as large as they did on the NWEA tests, but summer reading losses looked smaller on Renaissance tests than on NWEA tests.

There was just as little agreement on more complex questions concerning inequality in learning. Across tests, the gap between children in low-poverty and high-poverty schools grew faster during summer on some tests but not on others. Gaps between white and black, Asian, or Hispanic children displayed no consistent pattern of faster or slower growth during the summer. Across tests, boys seemed to learn math faster than girls during school, but the difference in school-year learning rates was small—one percent of a standard deviation per month or less—and the result did not replicate in reading. Moreover, there were many previous studies where no seasonal gender differences were observed.

Agreement did not improve when we raised our sights from the gaps between defined groups to the total variance of test scores. Across tests, there was no consistent evidence that skill variance grew faster during summer than during school.

Results did not just disagree on whether gaps and variance grew faster during summer or school—results even disagreed on whether gaps and variance grew at all as children grew older. If tests cannot agree on whether inequality grows over a period of years, how can we expect them to agree on whether inequality grows over the months of summer vacation?

Looking back at our results, it is not hard to find an intriguing pattern here or there, on one test or another. If all we had was one test, it would be tempting to weave a story around the most intriguing results. But the story seems much less persuasive if the results do not replicate in other data.

Why Do Modern Results Still Disagree with One Another?

Recent commentary has attributed nonreplicable summer learning results to practices common in older research, such as fixed-form tests and antiquated scaling methods (von Hippel 2019; von Hippel and Hamrock 2019). Yet the data used in this article used modern practices such as adaptive testing and IRT ability scaling. Even so, most results failed to replicate.

Other summer learning scholars have attributed replication failure to test dates that are too far from the beginning and end of summer vacation, or to inconsistent standards for classifying children into advantaged and disadvantaged groups (Alexander 2019; Heyns 1987). Yet we grouped children in the same way in every data set, and we adjusted for differences between test dates and the first and last day of summer vacation. Still, most results failed to replicate.

Some results in the summer learning literature are sensitive to model specification (Quinn 2015), but the disagreements among our results are not likely due to modeling. Disagreements were visible not just in the model output but in simple graphs of age trends in mean scores and score gaps.

What, then, might explain the disagreements between the data sources? A candidate explanation—initially attractive but inadequate on reflection—is that each data set sampled different children, and only the ECLS-K:2011 sample was nationally representative. Although the basic observation is correct, we doubt that sampling explains the discrepancies. Other NWEA studies, which weighted data to approximate national demographics, produced results very similar to ours, with large summer losses (Condron et al. 2021; Downey et al. 2022); in addition, our own auxiliary analyses of NWEA data found that summer losses looked large in nearly every district. In light of those findings, it seems unlikely that if the NWEA sample were nationally representative, it would show negligible summer losses like the ECLS-K:2011.

If we did insist on nationally representative samples, the implications for the summer learning literature would be devastating. In the long history of summer learning research, the vast majority of studies have used convenience samples, most of them local and much smaller than the samples used in this article. For example, the influential Beginning School Study enrolled 838 students in 20 Baltimore public schools. Only four summer learning studies have used nationally representative samples (the Sustaining Effects Study, Prospects, and the two ECLS-K cohort studies [1999 and 2011]), and only one nationally representative study, conducted more than 40 years ago, included any summers after second grade (the Sustaining Effects Study). Among nationally representative studies, none found that average skills declined over the summer, and none found that achievement gaps grew any faster during the summer than during the school year—if they grew at all.

In short, if we limited the literature to nationally representative samples, it would support hardly any popular claims about summer learning loss.

Another clearly inadequate explanation is that the data sources tested children in different grades. This isn't completely true. Although each source measured a different *range* of grades, there was considerable overlap, and when we focused on the grades that different sources had in common, agreement did not improve. For example, in the NWEA data the largest summer losses occurred in the summer after first grade—yet in the ECLS-K:2011 the losses in that summer were negligible.

So why did the results differ so much? After eliminating other explanations, we are left with test score measurement. Although all the tests in our study used IRT scaling, they did not use the same IRT model, nor did they use the same adaptive testing procedure. Moreover, IRT scaling does not, by itself, guarantee a vertical interval scale that can unambiguously compare children of different skill levels and ages (Briggs 2010). Within the IRT scaling framework, apparently small technical decisions can affect a score's properties (Bolt, Deng, and Lee 2014; Briggs and Weeks 2009) so that different IRT scores, all claiming to be vertically scaled, can give very different impressions of whether score gaps grow or shrink with age, as well as other issues like whether young children learn faster than older children (Bolt et al. 2014).

It is also possible that different tests measured somewhat different skills. "Math" and "reading" are broad categories, and Cooper et al. (1996) reported that different aspects of math and language skills were not equally susceptible to summer learning loss. For example, math computation decayed rapidly during the summer, but math concepts persisted reasonably well; likewise, spelling decayed rapidly, but vocabulary persisted or even grew (Cooper et al. 1996). Although the different assessments used in this article were aligned with state standards and the Common Core, that does not mean that they emphasized exactly the same skills and gave each skill exactly the same weight. If some math tests, for example, emphasized computation, whereas others emphasized concepts, we might expect different tests to show different amounts of summer learning loss—as they do.

The increasing reliance of summer learning research on restricted or proprietary test data—data in which we cannot see specific items or identify the skills that improve or decay over the summer—is a crucial limitation. Twenty-six years ago, Cooper et al.'s (1996) summer meta-analysis could break reading and math skills into components such as spelling and vocabulary, or computation and concepts. The fact that we can rarely make such fine-grained distinctions today and instead can only make broad and often nonreplicable statements about "math" or "reading" skills is a sign that the field has not progressed as much as we would like. Future summer learning research should use more open data that permits researchers to see more precisely what skills are being measured and how.

It is worth noting that hardly any of the assessments used in summer learning research are explicitly designed to evaluate learning over the summer. In our study, for example, tests were aligned with state standards (NWEA, Renaissance) or national standards (ECLS-K:2011), which changed with the start of each grade. Using these tests to estimate summer learning is a bit of an "off-label" use, and it may be that it is challenging for the same test to measure skills specific to each grade level and still measure how much skill is lost over the summer between one grade and the next. Prior articles speculated that adaptive testing might be able to measure both summer learning and grade level skills without compromise (von Hippel 2019; von Hippel and Hamrock 2019). But perhaps it is not that simple.

Most of the points we have just made about test scores are speculative. To verify them, we would need to drill down into the tests to see the content of each question, which students answered each question correctly and incorrectly, and how exactly scores were scaled. We would need to apply different scaling methods to the same tests and see how much difference they make. Unfortunately, we cannot do that because the vendor tests are proprietary and details about test used by the ECLS-K:2011 are highly restricted.[10] The increasingly reliance of summer learning research on inscrutable black-box tests will limit the insights that can emerge.

Another issue with vendor data is that they do not come from a designed study. Vendor data are dynamic; they are constantly being updated with new scores. In addition, vendors do not always have consistent, well-documented procedures for providing data in response to investigators' requests. Even when investigators get data from the same vendor, they often get different samples. This raises concerns about fairness if some researchers get larger or better samples than others. It also raises fundamental concerns about transparency and whether results obtained from vendor data can be reproduced.

Conclusion

What should we make of claims like "SES gaps grow only during summer" or "girls learn faster than boys during summer"? They are accurate descriptions of the data where they were observed, but they have limited generalizability. They do not replicate reliably in other data.

The limited replicability of summer learning results raises fundamental concerns because sociologists have long hoped that summer learning studies can illuminate the sources of inequality (Downey et al. 2004; Downey and Condron 2004; Hayes and Grether 1969; Heyns 1978). For example, when a gap or variance grows faster during summer, that has often been taken as evidence that inequality originates primarily outside school walls—not just in early childhood, before schooling begins, but every time school lets out for summer, and perhaps for shorter breaks (even weekends and evenings) as well. But if a result can only be obtained on one test, and fails to replicate on others, how much can it tell us about inequality in general?

We should emphasize again that summer learning programs still have potential even if it is unclear whether summer is a major source of inequality. One result that does replicate consistently is that advantaged children make little progress during summer—making summer a prime opportunity for disadvantaged children to catch up. Summer learning programs can have positive effects on student achievement. That said, attendance at summer programs can be poor, summer school effects are not necessarily larger for low-SES students, and many summers would be required to make a substantial dent in the achievement gap between low- and high-SES children, because the average effect of summer school (0.1 SD) is much smaller than the average achievement gap between high- and low-SES students ($>$ 1.0 SD).

Future researchers should shy away from drawing broad conclusions about summer or inequality from results obtained on a single test—no matter how large or representative the sample is. Instead, future research should compare results across multiple tests, consider measurement issues carefully, and only draw broad conclusions about results that replicate well.

Notes

[1] Standards for analysis and reporting were lower in the 1980s than they are today, and data from the Sustaining Effects Study are not readily available today. Ginsburg et al. (1981) was a conference paper, now available only on microfiche, and Klibanoff and Haggart (1988) was a technical report. Heyns' (1987) journal article was a commentary and reinterpretation of previous findings, rather than an independent analysis.

[2] Although the BSS started in 1982, the first results were not published until 1992, and those were limited to the first two years of the BSS. The 1992 publication was included in Cooper et al.'s (1996) meta-analysis, but later BSS publications, covering the later years of elementary school, were not.

[3] Each author analyzed restricted data from one of the data sets. The authors shared code and results but did not share restricted data with each other.

[4] The word "ability" is used here in its psychometric sense, where ability simply designates the student's current level of skill or knowledge. It should not be confused with the meaning of the term in economics and sociology, where ability may denote a trait that is fixed or innate.

[5] Renaissance assessments are called Star tests, and NWEA assessments are called MAP tests (measures of academic progress).

[6] To match the timing of the ECLS-K:2011, we requested Renaissance data from the fall of 2010, but we were told that Renaissance tests were undergoing recalibration at that time.

[7] In kindergarten, Renaissance gives an early literacy assessment, but it is less popular than the Renaissance reading assessments that start in grade 1.

[8] We identified these schools by running a chi-square test for uniformity on student frequencies by grade.

[9] Estimating a random effects model requires estimating not just the variance of each random effect but the covariances between them. However, the covariance between, say, the first and third grade learning rates cannot be estimated from an accelerated cohort design, because first and third grade learning cannot be observed for the same students. Software for random effects models did not catch this problem but did fail to converge because of it.

[10] Some tests publish subscores that appear to measure skill in different dimensions of reading and math skill. These subscores are not what they seem, however, because the IRT model assumes that all skills fit a single underlying dimension of "ability" (Koretz and Kim 2007).

References

Alexander, Karl L. 2019. "Summer Learning Loss Sure Is Real." *Education Next*. https://www.educationnext.org/summer-learning-loss-sure-is-real-response/.

Alexander, Karl L., Doris R. Entwisle, and Linda S. Olson. 2001. "Schools, Achievement, and Inequality: A Seasonal Perspective." *Educational Evaluation and Policy Analysis* 23(2):171–91. https://doi.org/10.3102/01623737023002171.

Atteberry, Allison, and Andrew McEachin. 2021. "School's Out: The Role of Summers in Understanding Achievement Disparities." *American Educational Research Journal* 58(2):239–82. https://doi.org/10.3102/0002831220937285.

Bolt, Daniel M., Sien Deng, and Sora Lee. 2014. "IRT Model Misspecification and Measurement of Growth in Vertical Scaling." *Journal of Educational Measurement* 51(2):141–62. https://doi.org/10.1111/jedm.12039.

Boulay, Beth, Barbara Goodson, Rob Olsen, Rachel McCormick, Catherine Darrow, Michael Frye, Katherine Gan, Eleanor Harvill, and Maureen Sarna. 2018. "The Investing in Innovation Fund: Summary of 67 Evaluations. Final Report." NCEE 2018-4013. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Briggs, Derek C. 2010. "Do Vertical Scales Lead to Sensible Growth Interpretations? Evidence from the Field." Presented at the 2010 Annual Meeting of the American Educational Research Association.

Briggs, Derek C., and Jonathan P. Weeks. 2009. "The Impact of Vertical Scaling Decisions on Growth Interpretations." *Educational Measurement: Issues and Practice* 28(4):3–14. https://doi.org/10.1111/j.1745-3992.2009.00158.x.

Burkam, David T., Douglas D. Ready, Valerie E. Lee, and Laura F. LoGerfo. 2004. "Social-Class Differences in Summer Learning between Kindergarten and First Grade: Model Specification and Estimation." *Sociology of Education* 77(1):1–31. https://doi.org/10.1177/003804070407700101.

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press. https://doi.org/10.1525/9780520969230.

Clemans, William V. 1993. "Item Response Theory, Vertical Scaling, and Something's Awry in the State of Test Mark." *Educational Assessment* 1(4):329–47. https://doi.org/10.1207/s15326977ea0104_3.

Clemans, William V. 1995. "Reply to Yen, Burket, and Fitzpatrick." *Educational Assessment* 3(2):191–202. https://doi.org/10.1207/s15326977ea0302_5.

Condron, Dennis J., Douglas B. Downey, and Megan Kuhfeld. 2021. "Schools as Refractors: Comparing Summertime and School-Year Skill Inequality Trajectories." *Sociology of Education* 94(4):316–40. https://doi.org/10.1177/00380407211041542.

Cooper, Harris, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-analytic Review." *Review of Educational Research* 66(3):227–68. https://doi.org/10.3102/00346543066003227.

DeMars, Christine. 2010. *Item Response Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001.

Downey, Douglas B., and Dennis J. Condron. 2004. "Playing Well with Others in Kindergarten: The Benefit of Siblings at Home." *Journal of Marriage and Family* 66(2):333–50.

Downey, Douglas B., and Dennis J. Condron. 2016. "Fifty Years since the Coleman Report: Rethinking the Relationship between Schools and Inequality." *Sociology of Education* 89(3):207–20. https://doi.org/10.1111/j.1741-3737.2004.00024.x.

Downey, Douglas B., Megan Kuhfeld, and Margriet van Hek. 2022. "Schools as a Relatively Standardizing Institution: The Case of Gender Gaps in Cognitive Skills." *Sociology of Education* 95(2):89–109. https://doi.org/10.1177/00380407211070319.

Downey, Douglas B., David M. Quinn, and Melissa Alcaraz. 2019. "The Distribution of School Quality: Do Schools Serving Mostly White and High-SES Children Produce the Most Learning?" *Sociology of Education* 92(4):386–403. https://doi.org/10.1177/0038040719870683.

Downey, Douglas B., Paul T. von Hippel, and Beckett A. Broh. 2004. "Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year." *American Sociological Review* 69(5):613–35. https://doi.org/10.1177/000312240406900501.

Duncan, Greg J., and Katherine Magnuson. 2011. "The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems." Pp. 47–70 in *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, edited by G. J. Duncan and R. J. Murnane. New York: Russell Sage Foundation.

Entwisle, Doris R., and Karl L. Alexander. 1992. "Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School." *American Sociological Review* 57(1):72–84. https://doi.org/10.2307/2096145.

Entwisle, Doris R., and Karl L. Alexander. 1994. "Winter Setback: The Racial Composition of Schools and Learning to Read." *American Sociological Review* 59(3):446–60. https://doi.org/10.2307/2095943.

Galbraith, Sally, Jack Bowden, and Adrian Mander. 2017. "Accelerated Longitudinal Designs: An Overview of Modelling, Power, Costs and Handling Missing Data." *Statistical Methods in Medical Research* 26(1):374–98. https://doi.org/10.1177/0962280214547150.

Gershon, Richard C. 2005. "Computer Adaptive Testing." *Journal of Applied Measurement* 6(1):109–27.

Gibbs, Benjamin G., and Douglas B. Downey. 2020. "The Black/White Skill Gap in Early Childhood: The Role of Parenting." *Sociological Perspectives* 63(4):525–51. https://doi.org/10.1177/0731121419896812.

Ginsburg, Alan, K. Baker, D. Sweet, and A. Rosenthal. 1981. "Summer Learning and the Effects of Schooling: A Replication of Heyns." Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA. Available on microfiche, https://eric.ed.gov/?id=ED204367.

Goldhaber, Dan, Lesley Lavery, and Roddy Theobald. 2015. "Uneven Playing Field? Assessing the Teacher Quality Gap between Advantaged and Disadvantaged Students." *Educational Researcher* 44(5):293–307. https://doi.org/10.3102/0013189X15592622.

Hayes, Donald P., and Judith Grether. 1969. "The School Year and Vacations: When Do Students Learn?" Presented at the Eastern Sociological Association Convention, New York, April 19, 1969.

Hayes, Donald P., and Judith Grether. 1983. "The School Year and Vacations: When Do Students Learn?" *Cornell Journal of Social Relations* 17(1):56–71.

Hecht, Amelie A., Keshia M. Pollack Porter, and Lindsey Turner. 2020. "Impact of the Community Eligibility Provision of the Healthy, Hunger-Free Kids Act on Student Nutrition, Behavior, and Academic Outcomes: 2011–2019." *American Journal of Public Health* 110(9):1405–10. https://doi.org/10.2105/AJPH.2020.305743.

Heckman, James J., and Dimitriy V. Masterov. 2007. "The Productivity Argument for Investing in Young Children." *Applied Economic Perspectives and Policy* 29(3):446–93. https://doi.org/10.3386/w13016.

Heyns, Barbara. 1978. *Summer Learning and the Effects of Schooling*. New York: Academic Press.

Heyns, Barbara. 1987. "Schooling and Cognitive Development: Is There a Season for Learning?" *Child Development* 58(5):1151–160. https://doi.org/10.2307/1130611.

Ioannidis, John P. 2005. "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research." *Journal of the American Medical Association* 294(2):218–28. https://doi.org/10.1001/jama.294.2.218.

Johnson, Angela, and Megan Kuhfeld. 2020. "Fall 2019 to Fall 2020 MAP Growth Attrition Analysis." Technical brief. Portland, OR: Northwest Evaluation Association (NWEA). https://www.nwea.org/research/publication/fall-2019-to-fall-2020-map-growth-attrition-analysis/.

Klibanoff, Leonard S., and Sue A. Haggart. 1981. "Summer Growth and the Effectiveness of Summer School." Technical report. System Development Corporation.

Koretz, Daniel, and Young-Suk Kim. 2007. "Changes in the Black–White Test Score Gap in the Elementary School Grades." CSE Report 715. Los Angeles, CA: Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kuhfeld, Megan. 2019. "Surprising New Evidence on Summer Learning Loss." *Phi Delta Kappan* 101(1):25–29. https://doi.org/10.1177/0031721719871560.

Kuhfeld, Megan, James Soland, Beth Tarasawa, Angela Johnson, Erik Ruzek, and Jing Liu. 2020. "Projecting the Potential Impact of COVID-19 School Closures on Academic Achievement." *Educational Researcher* 49(8):549–65. https://doi.org/10.3102/0013189X20965918.

Lee, Valerie E., and David T. Burkam. 2002. *Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School.* Washington, DC: Economic Policy Institute.

McCall, Martha S., G. Gage Kingsbury, and Allan Olson. 2004. "Individual Growth and School Success." Technical report. Lake Oswego, OR: Northwest Evaluation Association.

McEachin, Andrew, and Allison Atteberry. 2017. "The Impact of Summer Learning Loss on Measures of School Performance." *Education Finance and Policy* 12(4):468–91. https://doi.org/10.1162/edfp_a_00213.

National Center for Education Statistics. 2021. "Early Childhood Longitudinal Studies (ECLS) Program: Direct Cognitive Assessments." Retrieved December 30, 2021. https://nces.ed.gov/ecls/assessments2011.asp.

National Summer Learning Association (NSLA). 2017. "Summer by the Numbers." http://www.summerlearning.org/wp-content/uploads/2017/05/SummerByTheNumbers.pdf.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716. https://doi.org/10.1126/science.aac4716.

Phillips, Meredith, and Tiffani Chin. 2004. "How Families, Children, and Teachers Contribute to Summer Learning and Loss." Pp. 269–92 in *Summer Learning: Research, Policies, and Programs*, edited by G. D. Borman and M. Boulay. New York: Routledge.

Quinn, David M. 2015. "Kindergarten Black–White Test Score Gaps: Re-examining the Roles of Socioeconomic Status and School Quality with New Data." *Sociology of Education* 88(2):120–39. https://doi.org/10.1177/0038040715573027.

Quinn, David M., and Q. Tien Le. 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). https://doi.org/10.1177/2332858418819995.

Quinn, David M., and Morgan Polikoff. 2017. "Summer Learning Loss: What Is It, and What Can We Do about It?" Report. Washington, DC: Brookings Institution. https://www.brookings.edu/research/summer-learning-loss-what-is-it-and-what-can-we-do-about-it/.

Redding, Christopher. 2019. "A Teacher Like Me: A Review of the Effect of Student–Teacher Racial/Ethnic Matching on Teacher Perceptions of Students and Student Academic and Behavioral Outcomes." *Review of Educational Research* 89(4):499–535. https://doi.org/10.3102/0034654319853545.

Set, Andy. 2018. "Study Concludes MAP Growth Items Align to Common Core State Standards." NWEA Blog, February 27, 2018. https://www.nwea.org/blog/2018/study-concludes-map-growth-items-align-common-core-state-standards/.

Tourangeau, Karen, Christine Nord, Thanh Lê, Kathleen Wallner-Allen, Mary C. Hagedorn, John Leggitt, and Michelle Najarian. 2015. *User's Manual for the ECLS-K:2011 Kindergarten–Second Grade Data File and Electronic Codebook (NCES 2015-050)*. Washington, DC: US Department of Education.

von Hippel, Paul T. 2015. "Year-Round School Calendars: Effects on Summer Learning, Achievement, Parents, Teachers, and Property Values." Pp. 208–30 in *The Summer Slide: What We Know and What We Can Do about Summer Learning Loss*, edited by K. Alexander, S. Pitcock, and M. Boulay. New York: Teachers College Press.

von Hippel, Paul T. 2019. "Is Summer Learning Loss Real?" *Education Next* 19(4). https://www.educationnext.org/is-summer-learning-loss-real-how-i-lost-faith-education-research-results/.

von Hippel, Paul T., and Caitlin Hamrock. 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. https://doi.org/10.15195/v6.a3.

von Hippel, Paul T., Joseph Workman, and Douglas B. Downey. 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. https://doi.org/10.1177/0038040718801760.

Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods & Research* 23(2):230–57.

Yen, Wendy M., George R. Burket, and Anne R. Fitzpatrick. 1995a. "Rejoinder to Clemans." *Educational Assessment* 3(2):203–6. https://doi.org/10.1207/s15326977ea0302_6.

Yen, Wendy M., George R. Burket, and Anne R. Fitzpatrick. 1995b. "Response to Clemans." *Educational Assessment* 3(2):181–90. https://doi.org/10.1207/s15326977ea0302_4.

Acknowledgments: Replication code is available at https://osf.io/f4jrb/. Unfortunately, we cannot share vendor data from NWEA and Renaissance Learning, but the ECLS-K:2011 is available to any researcher with a restricted-data license.

Joseph Workman: University of Missouri, Kansas City. E-mail: workmanj@umkc.edu.

Paul T. von Hippel: University of Texas, Austin. E-mail: paulvonhippel@utexas.edu.

Joseph Merry: Furman University. E-mail: joseph.merry@furman.edu