ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-864

Estimating Learning When Test Scores Are Missing: The Problem and Two Solutions

Paul T. von Hippel
University of Texas, Austin

Longitudinal studies can produce biased estimates of learning if children miss tests. In an application to summer learning, we illustrate how missing test scores can create an illusion of large summer learning gaps when true gaps are close to zero. We demonstrate two methods that reduce bias by exploiting the correlations between missing and observed scores on fall and spring tests taken by the same child. One method uses those correlations to multiply impute missing scores. The other method models the correlations implicitly, using child-level random effects. Widespread adoption of these methods would improve the validity of summer learning studies and other longitudinal research in education.

VERSION: October 2023

Suggested citation: von Hippel, Paul T.. (2023). Estimating Learning When Test Scores Are Missing: The Problem and Two Solutions. (EdWorkingPaper: 23-864). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/07bv-by90

Estimating Learning When Test Scores Are Missing: The Problem and Two Solutions

Paul T. von Hippel
University of Texas, Austin

Abstract

Longitudinal studies can produce biased estimates of learning if children miss tests. In an application to summer learning, we illustrate how missing test scores can create an illusion of large summer learning gaps when true gaps are close to zero. We demonstrate two methods that reduce bias by exploiting the correlations between missing and observed scores on fall and spring tests taken by the same child. One method uses those correlations to multiply impute missing scores. The other method models the correlations implicitly, using child-level random effects. Widespread adoption of these methods would improve the validity of summer learning studies and other longitudinal research in education.

Note: Data and code used in this paper are available at https://osf.io/3qd76/

Estimating Learning When Test Scores Are Missing: The Problem and Two Solutions

Learning is a central topic in education, and estimating learning requires testing student's skills at two or more points in time. Yet the problem of estimating learning becomes challenging if some students miss tests. When tests are missed, the children who are tested on different occasions may not be comparable, and straightforward techniques—such as graphing changes in the mean score over time, subtracting pretest scores from posttest scores, or fitting ordinary linear regression models—may produce biased and misleading impressions of learning. Different investigators analyzing the same data can arrive at different conclusions about which students learn fastest and when gaps in skills between different students grow.

The problem of missing test scores is widespread in education research, particularly in longitudinal studies that try to test the same students repeatedly over a period of years. Test scores can be missing for a variety of reasons. Students drop out of studies, transfer out of tested schools, or simply don't show up or log in for tests. Schools switch from one test vendor to another. During the COVID-19 pandemic, so many children missed tests that some observers simply didn't believe estimates of learning loss until schools reopened and testing became more widespread.

Methods exist that can obtain unbiased estimates from longitudinal data with missing test scores. One of the best methods is multiple imputation, which fills in missing scores with a distribution of plausible values (Allison, 2002; Rubin, 1987; Schafer, 1997). Yet few education researchers are trained to use multiple imputation skillfully, and some researchers even regard imputation with suspicion ("making up data"). Othe researchers believe that multiple imputation is not worth the trouble because it "never makes a difference," or use it only with reluctance, as a "robustness check," hoping that it merely confirms the results they obtain from simpler methods—and unsure what to do if it doesn't.

Another method that can produce unbiased estimates from longitudinal data with missing test scores is maximum likelihood estimation with random effects at the student level and, when relevant, the school level as well (Molenberghs & Kenward, 2007). Random effects models are popular among education researchers trained in psychology, sociology, and schools of education, who typically describe random effects models as multilevel or hierarchical growth models (Raudenbush & Bryk, 2001; Singer & Willett, 2002), yet researchers trained in the econometric tradition often regard random effects with suspicion. When random effects models are used in educational research, missing values are never offered as a point in their favor.

In this article, we demonstrate that missing values can introduce serious biases into estimates of learning. Comparing the learning of students in public and private schools, we show that casual treatment of missing data can create an illusion that skill gaps between public and private students grow dramatically during summer vacation when in fact they grow very little, if they grow at all. The illusion of summer learning gaps is visible in simple trend graphs of mean test scores, and also shows up in regression estimates obtained by ordinary least squares.

We then show that the illusion dissipates when we use multiple imputation or random effects models, both of which suggest that the true summer learning gaps between public and private students are negligible. Yet not all learning gaps disappear when we use better methods for missing data. In particular, there is a persistent finding that public students gain on private students during kindergarten—a finding that deserves more thorough investigation.

The example of summer learning

The problems and solutions in this article should interest all researchers who are trying to estimate learning from longitudinal data. While the problems and solutions are general, our demonstration focuses on their application to summer learning. In recent years, several researchers have noticed that the results of summer learning studies are disturbingly "mixed" on nonreplicable. Most examples involve results obtained

from one dataset failing to replicate in another (Kuhfeld & Lewis, 2023; Quinn & Polikoff, 2017; von Hippel et al., 2018; von Hippel & Hamrock, 2019; Workman et al., 2023), but there are also examples where different analyses of the same or similar data can lead to different conclusions (Quinn, 2015). A variety of issues can make estimates of summer learning biased or non-replicable. Test dates are often far from the first and last day of summer (Burkam et al., 2004; Cooper et al., 1996; Downey et al., 2004; Heyns, 1987; Klibanoff & Haggart, 1981; Phillips & Chin, 2004); tests often change content or format over the summer (Reed et al., 2021; von Hippel & Hamrock, 2019); and the scaling of test scores can affect whether score gaps appear to grow or shrink as children get older (von Hippel, 2019; von Hippel et al., 2018; von Hippel & Hamrock, 2019). Even when analyzing the same data, researchers' choice of statistical model or operational definitions can affect whether score gaps appear to widen or narrow during summer or during school (Quinn, 2015; Quinn & McIntyre, 2017).

In this article, we show that missing data, too, can make summer learning estimates biased or hard to replicate. Missing data in summer learning studies is often substantial. For example, in the Early Childhood Longitudinal Studies of the Kindergarten classes of 1998-99 and 2010-11 (ECLS-K:1999 and ECLS-K:2011), only one-third of schools that were tested in the spring before summer vacation were also tested in the fall afterward. In summer learning studies that rely on tests given by vendors (such as Renaissance Learning or NWEA), about a quarter of children who have test scores one year do not take tests from the same vendor the next year (Johnson & Kuhfeld, 2020; Workman et al., 2023). Vendor tests can be missing for a variety of reasons: some students are absent during the testing window; others are exempted from tests or opt out; some students transfer to a school that does not use the same test vendor; others remain at the same school, but the school's vendor contract expires. In addition to test scores, student and school characteristics can be missing as well; for example, in test data collected by Renaissance Learning, student race and gender are optional, and many schools choose not to report them (Workman et al., 2023).

Like other longitudinal studies, summer learning studies often treat missing data casually, deleting missing values by default (a practice known as listwise deletion) without considering how deletion might

affect the estimates. There are a handful of summer learning studies that fill in missing values using multiple imputation (Alexander et al., 2007b, 2007a; Davies & Aurini, 2013; Downey et al., 2004, 2008; von Hippel & Hamrock, 2019), one study where the data were reduced to a subset with fewer missing values (Workman et al., 2023), and one study that investigated the differences between students with complete and incomplete data (Kuhfeld et al., 2021), but these are more the exception than the rule. In addition, some summer learning included random effects in the context of multilevel growth models (e.g., Borman & Dowling, 2006; Condron et al., 2021; Downey et al., 2004; Kuhfeld et al., 2021; Olson, 2004; von Hippel et al., 2018), but this was pure good fortune, as missing data was never mentioned as a justification for the model. In addition, there are a number of summer learning studies that did not include random effects (or use multiple imputation), even though they were quite careful with respect to other methodological issues (e.g., Davies et al., 2022; Gershenson & Hayes, 2017; Quinn, 2015; and every summer learning study conducted before 2004).

Our demonstration focuses on comparing school year and summer learning in public and private schools—a comparison that, while less popular than comparisons of children of different racial, ethnic, or socioeconomic groups, has received some attention in the summer learning literature (Carbonaro, 2003; Dallavis et al., 2021; Downey et al., 2008). The results speak to the substantive question of whether private schools do more or less than public schools to increase children's learning—a topic that has inspired considerable research, with mixed findings, since the 1980s (Coleman et al., 1982; Lubienski & Lubienski, 2013; Reardon et al., 2009; Shakeel et al., 2016). Separating summer and school learning sheds light on the differences between schools because learning differences during summer reflect family effects—that is, differences between the types of families who send their children to public or private school—while learning differences during the school year reflect the combined effects of families and schools. For example, if private students learn faster during summer but public students learn faster during school, then it would seem that the higher achievement of children in private schools is due to families, and public schools may actually be more effective at increasing student learning, at least on average.

We present the data, methods, and results in two passes. First, we ignore the problem of missing data and analyze the data using methods that would be appropriate if the data were complete. Then we acknowledge the problem of missing data and analyze the data using methods appropriate for incomplete data. The results, we will show, are strikingly different.

First Pass, Ignoring the Problem of Missing Data

We used data from the ECLS-K:2011, a study funded by the US Department of Education. The ECLS-K:2011 began in the 2010-11 school year with a nationally representative probability sample of 18,174 kindergartners in 796 public and 154 private schools. The ECLS-K:2011 was a two-stage cluster sample, in which children were sampled at random within schools, and schools were sampled at random within 90 primary sampling units, each of which was a single large county or a contiguous group of small counties. In keeping with the sample design, our analyses treat children as clustered within the schools where they were first sampled, even if they moved to a different school or left the school system later.

Children were tested in the fall and spring of kindergarten, first grade, and second grade. Scores on these six occasions enable us to estimate learning separately during each of the three school years and the two summers between. Children were also tested in the spring of fourth and fifth grade, but without fall tests in those years, there is no way to separate school year learning from summer learning.

The ECLS-K:2011 has been used in several previous studies of summer learning (e.g., Quinn et al., 2016; Quinn & Le, 2018; von Hippel et al., 2018; Workman et al., 2023). However, previous studies of summer learning in the ECLS-K:2011 have not compared public and private schools, as we do here.

Test Timing and Scaling

There are several issues that might affect estimates of summer learning in the ECLS-K:2011. One issue is the timing of tests. Tests were not given on the first and last day of school; instead, test dates varied from school to school, with average fall test dates in mid-October, nearly two months after school started,

and average spring test dates in late April, more than a month before school ended. What this means is that we cannot estimate summer learning just by subtracting spring tests from fall tests, because some of the time between spring and fall tests was spent in school. By comparing test dates to dates for the first and last day of school, we estimate each child's school exposure, and our models adjust for school exposure, as we explain later.

It is also important to describe the scaling of test scores, as scaling can affect summer learning estimates as well. Test scores were scaled using a 3-parameter logistic item response theory (IRT) model that estimated children's reading and math skill controlling for test items' difficulty, guessability, and sensitivity[1] (Tourangeau et al., 2017). Tests were scored on a "theta scale" that represented the child's log odds of correctly answering a question of reference difficulty, guessability, and sensitivity. On the theta scale, test scores could take positive or negative values, which can confuse some readers, but higher test scores still represent greater skill.

Descriptive results

Figure 1 plots the average of public and private students' observed reading and math scores on the fall and spring tests in kindergarten, first grade, and second grade. The dates on the horizontal axis are the average of the observed test dates.

It appears that private students pull away from public students during summer vacation. The opening of the gap between private and public students is evident in both subjects and both summers, but it is especially striking during the first summer in reading and the second summer in math. It also appears that public students gain on private students during each school year. The relative gains of public students

[1] In IRT, sensitivity refers to the steepness of the item response function—i.e., how quickly the probability of a correct answer rises as a function of student ability. Sensitivity can also be called "discrimination," but we avoid that term to avoid confusion with discrimination in the legal or sociological sense. Similar, a student's current skill level is often known in IRT as "ability," but we prefer "skill" because ability can connote skills that are fixed or innate.

appear especially large in first grade, where it appears that the gap between public and private students might be wiped out if the school year were approximately twice as long.

Model

A popular model in summer learning studies is the piecewise linear growth model, which equation (1) represents in both scalar and vector form:

$$\begin{aligned} Y_{cst} &= \alpha_0 + \gamma_0 Grade_{0,cst} + \beta_1 Summer_{1,cst} + \gamma_1 Grade_{1,cst} + \beta_2 Summer_{2,cst} + \gamma_2 Grade_{2,cst} + e_{cst} \\ &= \alpha_0 + \boldsymbol{\gamma}\, \mathbf{Grades}_{cst} + \boldsymbol{\beta}\, \mathbf{Summers}_{cst} + e_{cst} \end{aligned} \quad (1)$$

Here α_0 represents the starting point: the average score that would be obtained on a test given the first day of kindergarten. The vector $\boldsymbol{\gamma} = [\gamma_0\ \gamma_1\ \gamma_2]$ represents monthly learning rates during grades 0 (kindergarten), 1, 2, and 2, and the vector $\boldsymbol{\beta} = [\beta_1\ \beta_2]$ represent monthly learning rates during summer vacations 1 and 2. The model allows for different learning rates during each school year and summer, while assuming that learning within a given school year or summer proceeds at a constant rate—i.e., that scores increase in a linear fashion. By assuming linearity the model implicitly extrapolates beyond the observed test scores to the scores that would have been obtained on the first and last day of each school year.

The dependent variable Y_{cst} is the reading or math score of child *c* in school *s* on occasion *t=1,…6*. There are six test occasions, from the fall of kindergarten (*t*=1) to the spring of second grade (*t*=6). On each occasion, the $\mathbf{Grades}_{cst} = [Grade_{0,cst}, Grade_{1,cst}, Grade_{2,cst}]$ and $\mathbf{Summers}_{cst}=[Summer_{1,cst}, Summer_{2,cst}]$ vectors[2] contain variables representing the number of months that the child has spent in grade 0 (kindergarten), 1, and 2, and on their first and second summer vacations. For example, on test occasion *t*=3, in the fall of first grade, children have typically spent about $Grade_0$ =9.33 months in

[2] We adopt the convention that the elements of column vectors are separated by commas and the elements of row vectors are not. For example, $\mathbf{Summers}_{cst}=[Summer_{1,cst},\ Summer_{2,cst}]$ is a column vector, and $\boldsymbol{\beta} = [\beta_1\ \beta_2]$ is a row vector.

kindergarten, $Summer_1$ =2.67 months on summer vacation 1, $Grade_1$ =1.5 months in grade 1, and 0 months in grade 2 or summer vacation 2. Months of school and summer exposure are updated for each child on each test occasion.

To estimate differences between students in public and private students, we add a dummy variable for private schools and let it interact with the $\mathbf{Grades}_{\mathbf{cst}}$ and $\mathbf{Summers}_{\mathbf{cst}}$ variables:

$$Y_{cst} = \alpha_0 + \boldsymbol{\gamma}\, \boldsymbol{Grades}_{cst} + \boldsymbol{\beta}\, \boldsymbol{Summers}_{cst} + \Delta\alpha_0 Private_s + \Delta\boldsymbol{\gamma}\, Private_s \times \boldsymbol{Grades}_{cst} + \Delta\boldsymbol{\beta}\, Private_s \times \boldsymbol{Summers}_{cst} + e_{cst} \quad (2)$$

Here the coefficients $\Delta\alpha_0$, $\Delta\boldsymbol{\gamma}$, and $\Delta\boldsymbol{\beta}$ represent the difference between the average starting point and learning rates during in private and public schools. Differences between private and public learning rates can be different in each school year and summer. The error term e_{cst} is clustered by school to get cluster-robust standard errors (SEs).

We will call model (2) the OLS *analysis model*.

Observed-data OLS estimates with school-clustered SEs

Observed-data estimates from the OLS model appear in Table 1. In both reading and math, the average skill level at the start of kindergarten was significantly higher in private schools than in public schools (p<0.05). During the school years—kindergarten, grade 1, and grade 2—public school students shrink the gap by learning faster than private students; the difference between private and public students' learning is significant and favors public students in every school year. (p<.05). During summer vacations, however, the gap opens again as private students learn faster than public students. In reading, the summer difference is significant in both summers; in math, the difference is only significant in summer 2. Per month, the learning differences between public and private students are estimated to be larger in every summer than they are in any school year.

From these observed-data OLS estimates, it is tempting to conclude that public students learn faster than private students during the early school years, but private students learn faster than public students during the summers. If so, this would imply that public schools are more effective than private schools at raising test scores, at least in the early years, and that private students score higher only because of out-of-school advantages that are evident on the first day of kindergarten and during summer.

Second Pass, Acknowledging the Problem of Missing Data

The OLS estimates in the previous section would be unbiased if the data were complete, but in fact many test scores were missing. 78 percent of children missed at least one reading test, and one percent missed all six. As *p<0.05, **p<0.01, ***p<0.001. School-clustered standard errors in parentheses. The summer differences between public and private students are highlighted to draw attention.

Table 2 shows, the percentage of missing reading scores varied substantially across rounds and school types. In private schools, 3 percent of students missed tests in round 2, but 81 percent missed tests in round 3. Compared to private schools, public schools had higher rates of missing tests in kindergarten, but lower rates in first and second grade.

Many test scores were missing by design. One feature of the survey design was that recruitment of schools continued throughout the kindergarten year. In the fall of kindergarten, the sample included 16,464 children in 859 schools, but by the spring of kindergarten the sample included 18,174 children in 950 schools. That is, 10 percent of the spring kindergarten students were not in the fall kindergarten sample.

Another feature of this survey design was that fall testing in grades 1 and 2 was limited to a one-third random subsample of schools. This largely explains why over two-thirds of test scores were missing in the fall of grades 1 and 2. Some summer studies have limited these data to the one-third subsample, which, because it was random, was still representative of the kindergarten population in 2010-11.

Within the one-third subsample, rates of missingness were lower but still substantial in some rounds. For example, in round 5 (the fall of grade 2), the one-third subsample was missing scores for 21 percent of public students and 34 percent of private students.

In addition to scores that were missing by design, children missed test scores for a variety of reasons, known and unknown, that were beyond the study's control. For example, among the 1,031 students who missed the math test in the spring of kindergarten, 26 percent had moved to a county that was not part of the original sample, 9 percent could not be located, and 2 percent were being home-schooled. The remaining 64 percent were still attending sampled schools but missed the tests anyway, presumably because they were absent during the test window or refused to take the test, which was voluntary.

Listwise deletion

In our first pass, we handled missing data simply by dropping rounds with missing test scores from each child's data. For example, if a child was missing scores only in rounds 3 and 5, we dropped those rounds but retained the observed scores from rounds 1, 2, 4, and 6.

This approach is called *listwise deletion*, or *complete case analysis*. It is the most common approach to missing data, and the default in most software. But combining OLS with listwise deletion is not the best approach in general, and here it is strongly biased, as we will see.

Could scores be missing at random?

Missing test scores Y make estimates less precise, but do they also bias the OLS estimates? It depends. In data where the X variables are complete, as ours approximately[3] are, OLS regression estimates are unbiased if Y values are *missing at random* (MAR) (Little, 1992; von Hippel, 2007), meaning that the probability that a test score Y is missing is independent of the value of Y, controlling for X. Note that the probability that a test score Y is missing can depend on X; for example, more scores could be missing in some rounds than in others; or more scores in could be missing in public schools than in private schools.

[3] The private school dummy is complete, and the test date is observed whenever the test score is observed. Some schools are missing the start and end dates that are needed to calculate school and summer exposures, but we analyzed the data as though every school began on August 25 and ended on June 5. In other papers, we have imputed missing school dates, but it has never affected the estimates because school dates vary so little. In this paper, we treated school dates as a constant to keep the focus on missing test scores.

But the probability of a missing score cannot depend on the score itself cannot depend on the score itself once those X variables are controlled.

We cannot test the MAR assumption directly, for the simple reason that if a score Y is missing its value cannot be observed. In longitudinal data, however, we can test the plausibility of the MAR assumption indirectly by asking whether observed test scores in one round predict whether scores are missing in the next round. The serial correlation between scores from consecutive rounds is strong (between 0.75 and 0.89, depending on the subject and round), so if children with low scores in one round were more likely to miss tests in another round, it is almost sure that their scores in the missed round would also have been low if they were not missing.

Table 3 exploits this observation to test the MAR assumption indirectly. Using logistic regression, it predicts the probability of a missing score in round 3 (fall of grade 1) as a function of the score in round 2 (spring of kindergarten), a dummy for private schools, and an interaction between the private school dummy and the round 2 score.

The results suggest that test scores are not MAR, because round 2 scores predict whether scores are missing in round 3. The round 2 score has a small positive coefficient, meaning that, in public schools, children with *high* round 2 scores were more likely to miss round 3 tests. But the interaction between the round 2 score and the private school dummy has a larger negative coefficient, meaning that, in private schools, children with *low* round 2 scores were more likely to miss round 3 tests.

Think about what that means. In round 2, before summer vacation, nearly all children were tested, so average test scores were close to unbiased. But in round 3, after summer vacation, high-scoring children were more likely to miss tests in public schools, and low-scoring children were more likely to miss tests in private schools. In other words, the average score in private schools was biased high, and the average score in public schools was biased low. Therefore the round 3 gap between private and public schools was exaggerated (positively biased), and so was the growth in that gap over summer vacation. This bias or exaggeration affected the OLS estimates in Table 1.

Before moving on, we should make a couple of points about the logistic regression estimates in Table 3. First, although the estimates highlight the possibility of bias from missing data, missing data does not bias the results of the logistic regression itself—at least not much. In fact, the variables that are used in the logistic regression are nearly complete. The private school dummy is 100 percent complete, and the round 2 test scores are 94 percent complete. The round 3 test scores are incomplete, of course, but they are not in the model; instead, the model uses a dummy variable for whether the round 3 test score is missing, and that dummy variable is observed for every child, whether they have a round 3 test score or not. We could have fit similar logistic regression models using different rounds of test scores as predictors, but we chose to predict with round 2 test scores, because those have the fewest missing values.

Second, to conclude that *Y* values are not MAR, we must show that missingness depends on the *Y* values net of the *X* variables in the OLS regression. Does the logistic regression model control adequately for those X variables? It does. Like the OLS regression, the logistic regression includes a dummy for private schools. Like the OLS regression, the logistic regression involves the private school dummy in an interaction. The OLS regression includes school and summer exposures, which the logistic regression model does not include explicitly. But the logistic regression does control for exposures *implicitly*—by limiting analysis to rounds 2 and 3. Controlling for round effectively controls for exposure, because exposure varies little within rounds.

To sum up: net of the *X* variables in the OLS regression model, test scores *Y* are not MAR. Therefore the OLS regression estimates are biased, substantially overestimating the summer learning gaps between private and public schools.

How to make the MAR assumption more plausible

What can we do to improve estimates of school and summer learning? We must do something to increase the plausibility of the MAR assumption. Stated generally, MAR means that the probability a test score is missing does not depend on that test score, net of other variables *in the model*. As we have seen, the

MAR assumption is not plausible in the OLS regression model, which only conditions test scores Y on X variables such as school sector and school and summer exposures. But the MAR assumption would be more plausible if the model also conditioned test scores from one round *on test scores from other rounds.* As we have already pointed out, test scores from different rounds are strongly correlated, and scores from one round predict whether scores are missing from other rounds.

Estimates from incomplete data can improve dramatically when models include variables that strongly predict both whether values are missing and what the missing values are. Better prediction of missing values reduces the *standard error* of estimates, and better prediction of whether values are missing reduces *bias* by increasing the plausibility of the MAR assumption. When predictive variables that are not in the analysis model are pulled in just to help with missing values, they are called *auxiliary variables* (Collins et al., 2001; von Hippel & Lynch, 2013). In our situation, test scores from other rounds were already in the analysis model, but we will respecify the model to take better advantage of the correlations among test scores and the fact that test scores from one round predict whether test scores from another round are missing. We will refer to test scores used in this way as *quasi-auxiliary* variables.

Before respecifying our models, we should acknowledge that even skillful use of test scores as quasi-auxiliary variables cannot guarantee that the data will become MAR. Like other statistical assumptions (e.g., normality, linearity, or exogeneity), the MAR assumption is not always fully attainable, but our estimates can be expected to improve as we come closer to satisfying it. It remains possible that the probability a test score is missing will depend on that score, even after scores from other rounds are taken into account. But the dependency should be much weaker, and the resulting bias should be much smaller, than it would be if the model neglected the relationship among test scores from different rounds, as the OLS regression model did.

We used two approaches to improve estimates by using scores from different rounds.

Approach 1. Multiple imputation of wide data

Our first approach was to pre-process the incomplete data with an *imputation model* that estimated the relationship between scores from different rounds. Once the imputation model was estimated, we used it to impute missing scores with values drawn at random from the conditional distribution of the missing scores given the observed scores. After imputation, we used OLS to fit the imputed data to the same analysis model (2).

Imputation Model

To prepare data for imputation, we reshaped the data into a wide format, with one row per child and a column for each math score, reading score, and test date in each of 6 different rounds—18 columns in all (Table 4). Putting the data in wide format made it more convenient to estimate the correlations between scores on the tests taken by the same students in different rounds.

To estimate those correlations, we fit a multivariate normal model—a popular imputation model that assumes variables are normally distributed and the relationships among variables is linear (King et al., 2001; Schafer, 1997). The multivariate normal model is characterized by a mean vector $\boldsymbol{\mu}$, which contains the means of all the variables, and a covariance matrix $\boldsymbol{\Sigma}$, which contains the variances of all the variables and the covariances among pairs of variables. Note that this is equivalent to estimating the standard deviation of each variable and the correlations among pairs of variables, because the standard deviation is just the square root of the variance, and the correlation between two test score variables, say Y_1 and Y_2, is a function of their variances and covariances: $Corr(Y_1, Y_2) = Cov(Y_1, Y_2)/\sqrt{Var(Y_1)Var(Y_2)}$.

We fit the model separately to public and private schools, so that the mean and covariance matrix could differ across public and private schools. This made the imputation model *compatible* with the analysis model (Bartlett et al., 2015). Specifically, since the analysis model (2) allowed public and private schools to have different intercepts and slopes, it was important for the imputation to allow public and private schools to have different means and covariance matrices. Since the means and covariance matrices of a multivariate

normal model can be transformed into the intercepts and slopes of a regression model, if the regression parameters can vary across school sectors, then the means and covariance matrix must be allowed to vary as well (von Hippel, 2009).

Most imputation models are at least slightly misspecified, but the misspecification does not matter unless it causes the imputation model to be incompatible with the analysis model . For example, our multivariate normal model assumes that test scores are normally distributed, which they are not, but this should not cause bias since there is nothing in the analysis model that assumes normality (von Hippel, 2013).

We used *multiple imputation*, imputing and analyzing the data multiple times, then combining the results using specialized formulas to obtain valid and replicable inferences (Rubin, 1987). Specifically, we imputed the data 100 times, which according to a well-justified formula was more than enough to ensure that the estimated standard errors would likely change by less than 5 percent if the data were multiply imputed again (von Hippel, 2020). Following standard Bayesian practice (Rubin, 1987), in each imputation we drew the parameters of the imputation model at random from the posterior parameter distribution; then imputed missing values at random conditionally on the imputation model with the drawn parameters.

After imputation, we fit the OLS regression model (2) to each of the 100 imputed data sets. We then combined the estimates using formulas developed for multiply imputed data. The formulas averaged the 100 point estimates, and calculated standard errors by combining variation within and between the 100 imputed datasets (Rubin, 1987).

Results

The top row of Figure 2 graphs the mean of the imputed scores from the fall of kindergarten through the spring of second grade. Notice that the summer gaps that seemed so obvious in Figure 1, which used only observed scores, simply disappear when imputed scores are included in the averages. In the imputed data, there is no visible summer learning gap between private and public schools. Evidently the

summer learning gap apparent in the observed scores was due almost entirely to high-scoring private students and low-scoring public students missing the fall test.

OLS regression estimates from the multiply imputed data appear in the second column of Table 5 (reading) and Table 6 (math). The estimates confirm the visual impression we got from Figure 2. When OLS was applied to observed data, there appeared to be large and significant differences between the summer learning rates of private and public students (the estimates from Table 1 are repeated in the first column of Table 5 and Table 6). But when OLS is applied to the imputed data, those summer differences become non-significant—statistically and substantively indistinguishable from zero.

Although imputation made the learning differences between public and private students disappear during summer, during the kindergarten school year the imputed data still suggest that public students gained on private students. The rate at which public students gained on private students during kindergarten was similar whether the data were imputed or not. In later school years, by contrast, imputation substantially reduced the estimated learning difference between public and private students; in fact, in first grade math and second grade reading, these learning differences became statistically indistinguishable from zero.

Although it may seem surprising that imputation could make summer learning gaps disappear, we did know that gaps estimated from the observed data were biased, and we expected that the bias would be reduced or eliminated by imputation. There are two ways to think about how imputation reduced bias here. A theoretical explanation is that the imputation model, because it conditioned imputed test scores on observed test scores from other rounds, brought the data much closer to satisfying the MAR assumption. A more concrete explanation is that the imputation model filled in scores from the missing part of the test score distribution, so that the distribution was more representative after imputation than before. For example, in round 3 (the fall of kindergarten), the imputation model filled in the scores of low-scoring private students and high-scoring public students who missed the test. After missing scores were imputed, the distribution of scores became more representative, and artificial differences that resulted from observed scores being non-representative disappeared.

Approach 2. Maximum likelihood with random effects

A second approach to missing test scores is to add *random effects* to the analysis model. While random effects models, more often called hierarchical or multilevel growth models, are popular in the summer learning literature, it is not widely appreciated that random effects models, at least when they are fit using maximum likelihood, are relatively robust to missing values of the dependent variable (Molenberghs & Kenward, 2007). We will now illustrate the advantages of random effects.

Method

A two-level random-effects model can be written as

$$Y_{cst} = \alpha_0 + \boldsymbol{\gamma}\, \boldsymbol{Grades}_{cst} + \boldsymbol{\beta}\, \boldsymbol{Summers}_{cst} + \Delta\alpha_0 Private_s + \Delta\boldsymbol{\gamma}\, Private_s \times \boldsymbol{Grades}_{cst} + \Delta\boldsymbol{\beta}\, Private_s \times \boldsymbol{Summers}_{cst} + u_s + r_{cs} + v_{cst} \quad (3)$$

This is identical to our linear regression model (2), except that the random effects model splits the error term, which the linear regression represented with a single term e_{cst}, into three components: a random school effect u_s, a random child effect r_{cs}, and a test-level residual v_{cst}.

Why do random effects in the error term matter? Because they imply correlations among the test scores. And as we noted earlier, a model that incorporates correlations among the test scores will come much closer to making the data MAR.

While the imputation model modeled the correlations among test scores explicitly, the random effects model does so implicitly. The fact that all tests taken by the same child share the same random effect r_{cs} implies that tests taken by the same child in different rounds have a residual correlation equal to $Var(r_{\mathrm{cs}})/(Var(r_{cs}) + Var(e_{cst}))$. Likewise, the fact that each school had a characteristic random effect u_s implies that tests taken by children in the same school are correlated as well.

When random effects are estimated by maximum likelihood, the likelihood effectively integrates over the distribution of possible values for the missing scores. In large samples, is something like imputing the data an infinite number of times (Rubin, 1987; von Hippel & Bartlett, 2021; Wang & Robins, 1998).

Note that including school random effects is not the same as using OLS and clustering standard errors at the school level. Both approaches model the within-school correlations, but school-clustered standard errors only use those correlations to adjust the estimated standard errors around OLS estimates that were obtained as if the errors were uncorrelated. In random-effects models, the modeled correlations affect the point estimates as well as the estimated standard errors.

Results

Random effects estimates from listwise deleted data appear in the last two columns of Table 5 (reading) and Table 6 (math). Unlike the OLS estimates, the random effects estimates are very similar whether values are imputed or not. This confirms the intuition that maximum likelihood estimation of a random-effects model is like implicitly imputing the data an infinite number of times. Nothing is gained if the data are explicitly imputed as well.

During summer vacations, the random-effects estimates confirm that there is little or no learning difference between private and public students. Most of the summer differences are nonsignificant, and even when significant they are several times smaller than the differences estimated by applying OLS to listwise deleted data. During kindergarten, though, all models continue to suggest that public students gain on private students. The estimated kindergarten learning difference is statistically significant and similar in size whether random effects are included or not, and whether values are imputed or not.

Approach 3. Restricting to the one-third subsample

Although multiple imputation and maximum likelihood with random effects are the best general-purpose approaches to longitudinal data with missing values of the dependent variables, there is a third

option that is specific to the ECLS-K:2011. That option is to restrict the data to the one-third random subsample of schools where tests were given in rounds 3 and 5. While this sample is smaller and will produce less *precise* estimates, it is still a random sample of the population and would produce *unbiased* estimates if it were complete. The one-third subsample does have missing test scores, and we have already seen that they are not missing at random, but since a smaller fraction of scores are missing, any estimates obtained from it will be less sensitive to how missing data is handled.

The bottom of Figure 2 graphs the mean of observed test scores from the one-third subsample from the fall of kindergarten through the spring of second grade. The results look very much like those for the full sample with imputed data. Summer learning gaps between private and public schools are not visible for either subject (math or reading) or any summer vacation (summer 1 or summer 2).

Table 7 and Table 8 give reading and math estimates obtained from the one-third subsample. In general the results are very similar whether values are imputed or not, and whether parameters are estimated using random effects or OLS. This confirms our intuition that, because the one-third subsample has fewer missing values than the full sample, estimates from the one-third subsample will be less sensitive to the treatment of missing data.

Conclusion

Our results show that estimates of summer learning can be sensitive to missing test scores. When we analyzed the data in the easiest and most casual way—simply dropping missing test scores and calculating means and OLS regressions from the observed values only—it appeared that there were substantial summer learning gaps between private and public schools. But when we treated the incomplete data using more appropriate methods—multiple imputation or maximum likelihood with random effects—the summer learning gaps shrank dramatically, in most cases to nothing.

This finding raises concerns about past summer learning studies that did not use multiple imputation or random effects, and it suggests that future summer learning studies should be alert to the challenges

posed by missing values. Future summer learning studies should describe their missing value problems carefully and handle it using either multiple imputation or random-effects models, or both.

Fortunately, random effects models are already popular in longitudinal studies of learning, though many studies do not use random effects, and those that do have never previously cited the advantages of random effects models for incomplete data. That said, random effects models do have limitations. First, random effects models only address missing values in the dependent variable (test scores). If a substantial number of values are also missing from the independent variables (e.g., race, SES, gender), then they will need to be filled in using multiple imputation.

Another issue with random effects models is that, unlike OLS, random effects estimates of learning rates do not represent simple changes in the mean score. Instead, random effects estimates estimates represent a weighted average of within- and between-school variation, with the within-school variation dominating the estimates if the number of children sampled per school is large (Greene, 2011; Wooldridge, 2002). In our study, only between-school variation contributed to the point estimates, because the X variable of interest was school sector, which varied only between schools. But in another study, comparing the learning of children whose characteristics vary both within and between schools (such as children of different ethnicities or family SES), within-school variation will dominate random effects estimates, and researchers should consider carefully whether within-school variation is what interests them. Researchers who are also interested in between-school variation will need to estimate that explicitly. A common recommendation is to fit a *within-between* model that includes both school-level averages of child characteristics and individual child characteristics, perhaps centered around their school-level averages (Allison, 2009, 2017; Raudenbush & Bryk, 2001).

There are some costs associated with using multilevel random effects models or multiple imputation. Multiple imputation increases runtime, both because the data need to be imputed and because the data then need to be analyzed multiple times. There are ways to make the process faster, but they are not widely implemented (von Hippel et al., 2023). Investigators using multiple imputation must also make sure that the

imputation and analysis models are compatible. In longitudinal studies, this typically means imputing test scores in wide format, and imputing separately any subgroups (e.g., public vs. private students) whose learning rates will be compared in the analysis model.

Multilevel random effects models can also run slowly and have convergence trouble, but this depends on the package. HLM and MPlus run these models much faster than SAS's MIXED procedure, R's *lme4* package, or especially Stata's *mixed* command (McCoach et al., 2018). If you like the missing-data properties of random effects models, you may wish to use a different package than you ordinarily do. The runtime issues are real, and more effort should be spent on implementing faster, more robust algorithm in packages that do not already use them.

But the extra minutes or hours needed to treat missing data properly pale in comparison to the years of readers' time wasted by publishing seriously misleading results that are biased by casual treatment of missing data. Accurate estimates are worth waiting for.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007a). Lasting Consequences of the Summer Learning Gap. *American Sociological Review*, *72*(2), 167–180. https://doi.org/10.1177/000312240707200202

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007b). Summer learning and its implications: Insights from the Beginning School Study. *New Directions for Youth Development*, *2007*(114), 11–32. https://doi.org/10.1002/yd.210

Allison, P. D. (2002). *Missing Data*. Sage Publications.

Allison, P. D. (2009). *Fixed Effects Regression Models* (1st ed.). Sage Publications, Inc.

Allison, P. D. (2017, July 26). Using "Between-Within" Models to Estimate Contextual Effects. *Statistical Horizons*. https://statisticalhorizons.com/between-within-contextual-effects/

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348

Borman, G. D., & Dowling, N. M. (2006). Longitudinal Achievement Effects of Multiyear Summer School: Evidence From the Teach Baltimore Randomized Field Trial. *Educational Evaluation and Policy Analysis*, *28*(1), 25–48. https://doi.org/10.3102/01623737028001025

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

Carbonaro, W. (2003). Sector Differences in Student Learning: Differences in Achievement Gains Across School Years and During the Summer. *Catholic Education: A Journal of Inquiry and Practice*, *7*(2), Article 2. https://doi.org/10.15365/joce.0702062013

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). *High school achievement: Public, Catholic, and private schools compared*. Basic Books. http://www.getcited.org/pub/102154890

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Condron, D. J., Downey, D. B., & Kuhfeld, M. (2021). Schools as Refractors: Comparing Summertime and School-Year Skill Inequality Trajectories. *Sociology of Education*, *94*(4), 316–340. https://doi.org/10.1177/00380407211041542

Cooper, H. M., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, *66*(3), 227–268. https://doi.org/10.3102/00346543066003227

Dallavis, J. W., Kuhfeld, M., Tarasawa, B., & Ponisciak, S. (2021). Achievement Growth in K-8 Catholic Schools Using NWEA Data. *Journal of Catholic Education*, *24*(2), 1–19. https://doi.org/10.15365/joce.2402012021

Davies, S., & Aurini, J. (2013). Summer Learning Inequality in Ontario. *Canadian Public Policy*, *39*(2), 287–307. https://doi.org/10.3138/CPP.39.2.287

Davies, S., Aurini, J., & Hillier, C. (2022). Reproducing or Reducing Inequality? The Case of Summer Learning Programs. *Canadian Journal of Education / Revue Canadienne de l'éducation*, *45*(4), 1055–1083. https://doi.org/10.53967/cje-rce.5311

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year. *American Sociological Review*, *69*(5), 613.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are "Failing" School Really Failing? Using Seasonal Comparisons to Evaluate School Effectiveness. *Sociology of Education*, *81*(3), 242–270.

Gershenson, S., & Hayes, M. S. (2017). The Summer Learning of Exceptional Students. *American Journal of Education*, *123*(3), 447–473. https://doi.org/10.1086/691226

Greene, W. H. (2011). *Econometric Analysis* (7th ed.). Prentice Hall.

Heyns, B. (1987). Schooling and Cognitive Development: Is There a Season for Learning? *Child Development*, *58*(5), 1151–1160.

Johnson, D., & Kuhfeld, M. (2020, November). Fall 2019 to fall 2020 MAP Growth attrition analysis. *NWEA*. https://www.nwea.org/research/publication/fall-2019-to-fall-2020-map-growth-attrition-analysis/

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, *95*(01), 49–69. https://doi.org/null

Klibanoff, L. S., & Haggart, S. A. (1981). *Summer growth and the effectiveness of summer school* (8; Study of the Sustaining Effects of Compensatory Education on Basic Skills.). RMC Research Corporation. http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED213789

Kuhfeld, M., Condron, D. J., & Downey, D. B. (2021). When Does Inequality Grow? A Seasonal Analysis of Racial/Ethnic Disparities in Learning From Kindergarten Through Eighth Grade. *Educational Researcher*, *50*(4), 225–238. https://doi.org/10.3102/0013189X20977854

Kuhfeld, M., & Lewis, K. (2023, January 30). *Is summer learning loss real, and does it widen test score gaps by family income?* Brookings. https://www.brookings.edu/articles/is-summer-learning-loss-real-and-does-it-widen-test-score-gaps-by-family-income/

Little, R. J. A. (1992). Regression With Missing X's: A Review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Lubienski, C. A., & Lubienski, S. T. (2013). *The Public School Advantage: Why Public Schools Outperform Private Schools* (Illustrated edition). University of Chicago Press.

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A. J., & Bellara, A. (2018). Does the Package Matter? A Comparison of Five Common Multilevel Modeling Software Packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102/1076998618776348

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. J. Wiley.

Olson, K. L. A., Doris R. Entwisle, Linda S. (2004). Schools, Achievement, and Inequality: A Seasonal Perspective. In *Summer Learning*. Routledge.

Phillips, M., & Chin, T. (2004). School Inequality: What Do We Know? In K. Neckerman (Ed.), *Social Inequality* (pp. 467–519). Russell Sage Foundation.

Quinn, D. M. (2015). Black–White Summer Learning Gaps Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis*, *37*(1), 50–69. https://doi.org/10.3102/0162373714534522

Quinn, D. M., Cooc, N., McIntyre, J., & Gomez, C. J. (2016). Seasonal Dynamics of Academic Achievement Inequality by Socioeconomic Status and Race/Ethnicity: Updating and Extending Past Research With New National Data. *Educational Researcher*, *45*(8), 443–453. https://doi.org/10.3102/0013189X16677965

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

Quinn, D. M., & McIntyre, J. (2017). Do learning rates differ by race/ethnicity over kindergarten? Reconciling results across gain score, first-difference, and random effects models. *Economics of Education Review*, *59*, 81–86. https://doi.org/10.1016/j.econedurev.2017.06.006

Quinn, D. M., & Polikoff, M. (2017). *Summer learning loss: What is it, and what can we do about it?* Brookings. https://www.brookings.edu/articles/summer-learning-loss-what-is-it-and-what-can-we-do-about-it/

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Sage Publications, Inc.

Reardon, S. F., Cheadle, J. E., & Robinson, J. P. (2009). The Effect of Catholic Schooling on Math and Reading Development in Kindergarten Through Fifth Grade. *Journal of Research on Educational Effectiveness*, *2*(1), 45–87. https://doi.org/10.1080/19345740802539267

Reed, D. K., Aloe, A. M., Park, S., & Reeger, A. J. (2021). Exploring the summer reading effect through visual analysis of multiple datasets. *Journal of Research in Reading*, *44*(3), 597–616. https://doi.org/10.1111/1467-9817.12357

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.

Shakeel, M. D., Anderson, K. P., & Wolf, P. J. (2016). *The Participant Effects of Private School Vouchers Across the Globe: A Meta-Analytic and Systematic Review* (SSRN Scholarly Paper ID 2777633). Social Science Research Network. https://papers.ssrn.com/abstract=2777633

Singer, J. D., & Willett, J. B. (2002). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Tourangeau, K., Nord, C., Thanh Lê, Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., Najarian, M., & Mulligan, G. M. (2017). *User's manual for the ECLS-K:2011 Kindergarten-second grade data file and electronic codebook, public version* (NCES 2017-285). U.S. Department of Education, National Center for Education Statistics.

von Hippel, P. T. (2007). Regression With Missing Ys: An Improved Strategy For Analyzing Multiply Imputed Data. *Sociological Methodology*, *37*, 83–117.

von Hippel, P. T. (2009). How To Impute Interactions, Squares, and Other Transformed Variables. *Sociological Methodology*, *39*(1), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

von Hippel, P. T. (2013). Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociological Methods & Research*, *42*(1), 105–138.

von Hippel, P. T. (2020). How Many Imputations Do You Need? A Two-Stage Calculation Using a Quadratic Rule. *Sociological Methods & Research*, *49*(3), 699–718. https://doi.org/10.1177/0049124117747303

von Hippel, P. T. (2019, June 4). *Is Summer Learning Loss Real? How I lost faith in one of education research's classic results*. Education Next. https://www.educationnext.org/is-summer-learning-loss-real-how-i-lost-faith-education-research-results/

von Hippel, P. T., Allison, P. D., & Williams, R. (2023). *How To Make Imputation Faster* [Unpublished manuscript].

von Hippel, P. T., & Bartlett, J. W. (2021). Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws. *Statistical Science*, *36*(3), 400–420. https://doi.org/10.1214/20-STS793

von Hippel, P. T., & Hamrock, C. (2019). Do Test Score Gaps Grow Before, During, or between the School Years? Measurement Artifacts and What We Can Know in Spite of Them. *Sociological Science*, *6*(3). http://dx.doi.org/10.15195/v6.a3

von Hippel, P. T., & Lynch, J. (2013). Efficiency Gains from Using Auxiliary Variables in Imputation. *arXiv:1311.5249 [Stat]*. http://arxiv.org/abs/1311.5249

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in Reading and Math Skills Forms Mainly before Kindergarten: A Replication, and Partial Correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

Wang, N., & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, *85*(4), 935–948. https://doi.org/10.1093/biomet/85.4.935

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data* (1st ed.). The MIT Press.

Workman, J., von Hippel, P. T., & Merry, J. (2023). Findings on Summer Learning Loss Often Fail to Replicate, Even in Recent Data. *Sociological Science*, *10*, 251–285. https://doi.org/10.15195/v10.a8

Figures

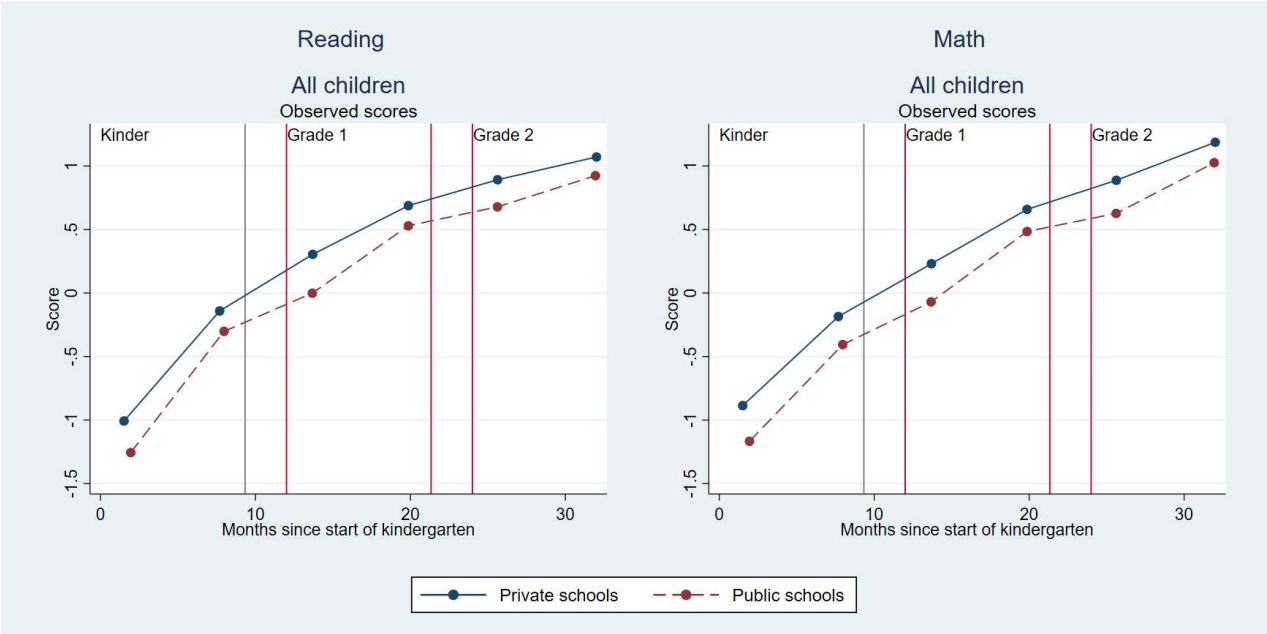

Figure 1. The average of observed scores in public and private schools in the fall and spring of kindergarten, first grade, and second grade.

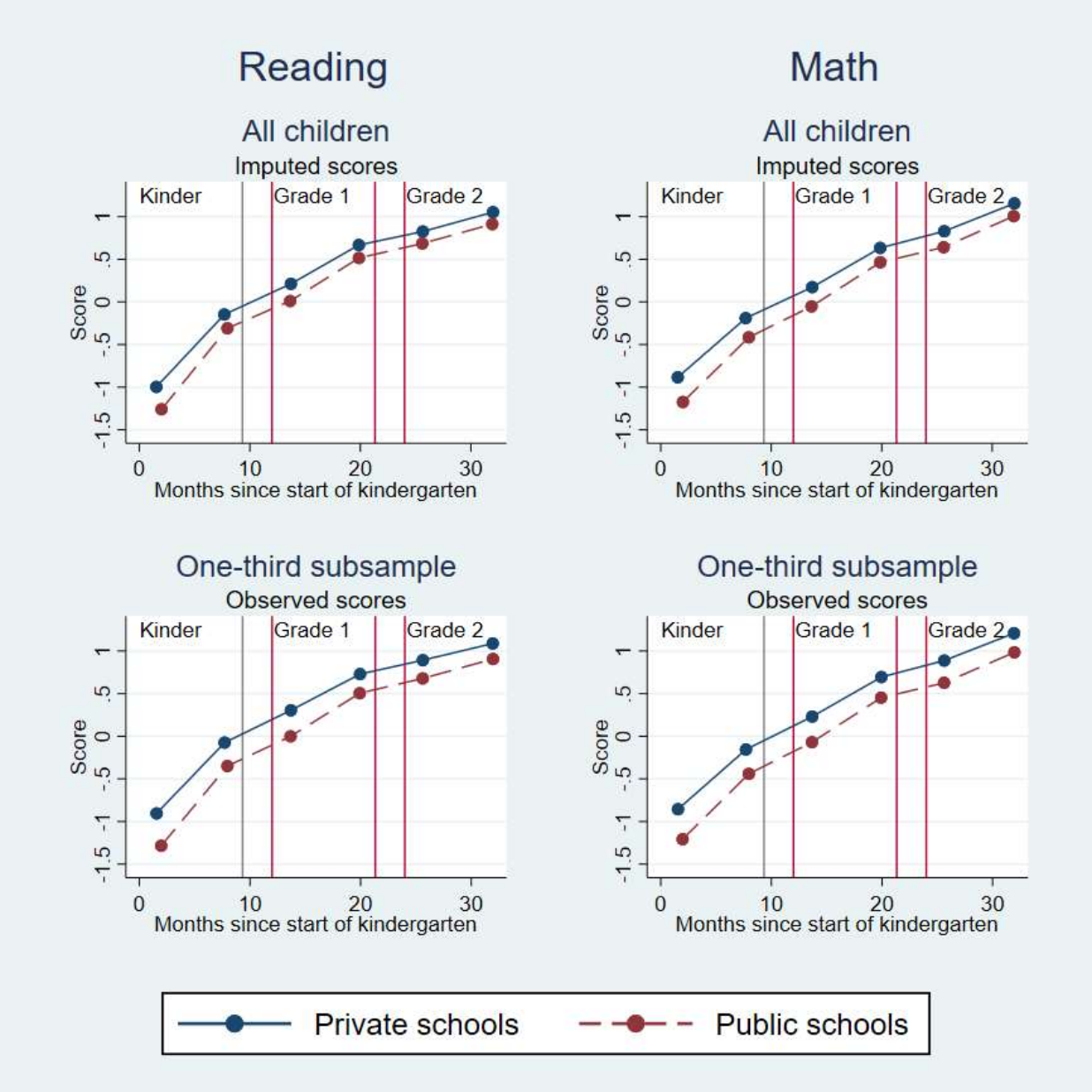

Figure 2. Gaps between private and public schools—using imputed data (top) and using a one-third subsample where fewer values are missing (bottom).

Tables

Table 1. OLS regression estimates obtained from listwise deleted data with observed test scores only.

	Reading	Math estimates
Public schools		
Score, start of kindergarten	-1.561***	-1.410***
	(0.016)	(0.014)
Monthly learning rates		
Kindergarten	0.158***	0.126***
	(0.002)	(0.001)
Summer 1	-0.018**	0.012
	(0.007)	(0.007)
Grade 1	0.084***	0.087***
	(0.002)	(0.002)
Summer 2	-0.011*	-0.028***
	(0.005)	(0.007)
Grade 2	0.038***	0.062***
	(0.001)	(0.002)
Difference between private and public schools		
Score, start of kindergarten	0.343***	0.350***
	(0.041)	(0.031)
Monthly learning rates		
Kindergarten	-0.018***	-0.012***
	(0.004)	(0.003)
Summer 1	0.059**	0.027
	(0.019)	(0.020)
Grade 1	-0.022***	-0.016*
	(0.006)	(0.006)
Summer 2	0.034*	0.046*
	(0.015)	(0.019)
Grade 2	-0.009*	-0.015**
	(0.004)	(0.005)
Scores	71,726	71,622

*p<0.05, **p<0.01, ***p<0.001. School-clustered standard errors in parentheses. The summer differences between public and private students are highlighted to draw attention.

Table 2. Percentage of reading scores that were missing.

	Full sample		One-third subsample	
Round	Public	Private	Public	Private
1. Fall kindergarten	15%	6%	17%	4%
2. Spring kindergarten	6%	3%	6%	4%
3. Fall 1st grade	71%	78%	14%	24%
4. Spring 1st grade	16%	23%	12%	16%
5. Fall 2nd grade	73%	81%	21%	34%
6. Spring 2nd grade	23%	31%	16%	21%
Children	15,953	2,221	5,458	651
Schools	796	154	275	47

Note. The percentage of math scores that were missing was practically identical.

Table 3. Logistic regression predicting which scores were missing in fall grade 1

	Missing score, fall grade 1	
Predictors	Reading	Math
Score, spring kindergarten	0.09+	0.11*
	(0.05)	(0.05)
Private school	0.32	0.32
	(0.21)	(0.21)
Private school × Score, spring kindergarten	-0.50**	-0.36+
	(0.18)	(0.19)
Constant	0.85***	0.86***
	(0.07)	(0.08)
Observations	17,186	17,143

*** p<0.001, ** p<0.01, * p<0.05, + p<0.10. School-clustered standard errors in parentheses.
Note. The model used spring scores to predict which fall scores were missing from the same subject. For example, reading scores in spring kindergarten were used to predict which students missed reading tests in fall grade 1.

Table 4. Twelve children's incomplete test data in wide format

	Round 1			Round 2			Round 3			Round 4			Round 5			Round 6		
Child ID	Date	Reading	Math	Date	Reading	Math	Date	Reading	Math	Date	Reading	Math	Date	Reading	Math	Date	Reading	Math
10000034	12/4/2010	-0.67	-0.79	5/4/2011	-0.08	-0.28	11/11/2011	0.37	0.38	5/11/2012	0.67	0.30	11/15/2012	0.71	0.49	5/15/2013	0.83	0.78
10000035	10/19/2010	-2.11	-2.78	4/27/2011	-1.87	-2.40												
10000036	12/4/2010	-0.53	-0.58	5/4/2011	-0.42	-0.10	10/19/2011	0.06	-0.01	5/4/2012	0.64	0.19	10/15/2012	0.84	0.39	4/15/2013	0.93	0.57
10000037	11/4/2010	-2.10	-2.11	4/11/2011	-1.86	-0.94	9/11/2011	-1.30	-0.85									
10000038	10/4/2010	-1.89	-0.72	4/4/2011	-0.66	-0.09				3/27/2012	0.26	0.55				5/15/2013	0.77	1.21
10000039	12/19/2010	-2.60		5/27/2011	-1.04	-0.74				5/11/2012	-0.18	-0.54						
10000040	9/11/2010	-1.88	-1.51	4/11/2011	-0.78	-0.44	9/19/2011	-0.05	0.17	3/27/2012	0.53	0.50	9/15/2012	0.89	0.87	3/15/2013	0.91	1.36
10000041	12/4/2010	0.05	0.14	5/11/2011	0.89	0.74				5/11/2012	0.96	1.44						
10000042	10/4/2010	-2.01	-1.82	4/4/2011	-0.33	-1.04				3/27/2012	0.18	-0.12				4/15/2013	0.59	0.35
10000043	10/19/2010	-1.92	-2.27	4/4/2011	-0.23	-0.71	10/11/2011	0.00	-0.53	4/19/2012	0.59	0.18						

Note. Each row contains observed and missing values from 6 rounds of tests on an individual child. These twelve rows are just an aid to visualization. In the actual data, there are 18,174 rows and each column has a different name.

Table 5. Reading results, all children

	OLS regression		Random effects regression	
	Observed scores	Imputed scores	Observed scores	Imputed scores
Public schools				
Score, start of kindergarten	-1.561***	-1.576***	-1.577***	-1.576***
	(0.016)	(0.015)	(0.010)	(0.009)
Monthly learning rates				
Kindergarten	0.158***	0.159***	0.160***	0.159***
	(0.002)	(0.001)	(0.001)	(0.001)
Summer 1	-0.018**	-0.010**	-0.014***	-0.013***
	(0.007)	(0.003)	(0.003)	(0.002)
Grade 1	0.084***	0.080***	0.081***	0.082***
	(0.002)	(0.001)	(0.001)	(0.001)
Summer 2	-0.011*	-0.000	0.002	-0.003
	(0.005)	(0.002)	(0.003)	(0.002)
Grade 2	0.038***	0.036***	0.034***	0.037***
	(0.001)	(0.001)	(0.001)	(0.001)
Difference between private and public schools				
Score, start of kindergarten	0.343***	0.367***	0.329***	0.341***
	(0.041)	(0.040)	(0.025)	(0.023)
Monthly learning rates				
Kindergarten	-0.018***	-0.021***	-0.020***	-0.021***
	(0.004)	(0.003)	(0.002)	(0.002)
Summer 1	0.059**	0.010	0.018*	0.015*
	(0.019)	(0.007)	(0.009)	(0.006)
Grade 1	-0.022***	-0.005*	-0.010***	-0.008***
	(0.006)	(0.002)	(0.003)	(0.002)
Summer 2	0.034*	-0.007	-0.010	0.002
	(0.015)	(0.005)	(0.009)	(0.006)
Grade 2	-0.009*	0.001	0.002	-0.002
	(0.004)	(0.001)	(0.003)	(0.002)
Scores	71,726	109,044	71,726	109,044

*p<0.05, **p<0.01, ***p<0.001. School-clustered standard errors in parentheses. The summer differences between public and private students are highlighted to draw attention.

Table 6. Math results, all children

	OLS regression		Random effects regression	
	Observed scores	Imputed scores	Observed scores	Imputed scores
Public schools				
Score, start of kindergarten	-1.410***	-1.424***	-1.422***	-1.424***
	(0.014)	(0.013)	(0.010)	(0.009)
Monthly learning rates				
Kindergarten	0.126***	0.126***	0.127***	0.127***
	(0.001)	(0.001)	(0.001)	(0.001)
Summer 1	0.012	0.023***	0.020***	0.019***
	(0.007)	(0.003)	(0.002)	(0.002)
Grade 1	0.087***	0.082***	0.083***	0.084***
	(0.002)	(0.001)	(0.001)	(0.001)
Summer 2	-0.028***	-0.011***	-0.012***	-0.015***
	(0.007)	(0.002)	(0.002)	(0.002)
Grade 2	0.062***	0.058***	0.057***	0.058***
	(0.002)	(0.001)	(0.001)	(0.001)
Difference between private and public schools				
Score, start of kindergarten	0.350***	0.365***	0.334***	0.342***
	(0.031)	(0.029)	(0.026)	(0.025)
Monthly learning rates				
Kindergarten	-0.012***	-0.013***	-0.014***	-0.014***
	(0.003)	(0.003)	(0.001)	(0.001)
Summer 1	0.027	-0.009	-0.002	-0.000
	(0.020)	(0.007)	(0.007)	(0.006)
Grade 1	-0.016*	-0.005	-0.008***	-0.009***
	(0.006)	(0.002)	(0.002)	(0.002)
Summer 2	0.046*	0.007	0.014	0.018**
	(0.019)	(0.006)	(0.008)	(0.006)
Grade 2	-0.015**	-0.006**	-0.007**	-0.008***
	(0.005)	(0.002)	(0.002)	(0.002)
Scores	71,622	109,044	71,622	109,044

*p<0.05, **p<0.01, ***p<0.001. School-clustered standard errors in parentheses. The summer differences between public and private students are highlighted to draw attention.

Table 7. Reading results, one-third subsample

	OLS regression		Random effects regression	
	Observed scores	Imputed scores	Observed scores	Imputed scores
Public schools				
Score, start of kindergarten	-1.591***	-1.608***	-1.604***	-1.607***
	(0.028)	(0.024)	(0.016)	(0.016)
Monthly learning rates				
Kindergarten	0.156***	0.157***	0.157***	0.158***
	(0.003)	(0.003)	(0.001)	(0.001)
Summer 1	0.000	-0.006	-0.006*	-0.008**
	(0.007)	(0.006)	(0.003)	(0.003)
Grade 1	0.080***	0.081***	0.082***	0.082***
	(0.001)	(0.001)	(0.001)	(0.001)
Summer 2	0.002	0.001	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)
Grade 2	0.036***	0.036***	0.037***	0.037***
	(0.001)	(0.001)	(0.001)	(0.001)
Difference between private and public schools				
Score, start of kindergarten	0.470***	0.500***	0.438***	0.455***
	(0.065)	(0.063)	(0.044)	(0.043)
Monthly learning rates				
Kindergarten	-0.020**	-0.024***	-0.023***	-0.024***
	(0.006)	(0.005)	(0.003)	(0.003)
Summer 1	0.015	0.007	0.011	0.010
	(0.014)	(0.011)	(0.009)	(0.009)
Grade 1	-0.013*	-0.010*	-0.012***	-0.011***
	(0.005)	(0.004)	(0.003)	(0.003)
Summer 2	0.004	-0.006	-0.002	0.000
	(0.010)	(0.007)	(0.010)	(0.009)
Grade 2	-0.004	-0.001	-0.002	-0.003
	(0.003)	(0.002)	(0.003)	(0.003)
Scores	31,261	36,654	31,261	36,654

*p<0.05, **p<0.01, ***p<0.001. School-clustered standard errors in parentheses. The summer differences between public and private students are highlighted to draw attention.

Table 8. Math results, one-third subsample

	OLS regression		Random effects regression	
	Observed scores	Imputed scores	Observed scores	Imputed scores
Public schools				
Score, start of kindergarten	-1.455***	-1.466***	-1.465***	-1.468***
	(0.025)	(0.022)	(0.017)	(0.016)
Monthly learning rates				
Kindergarten	0.127***	0.127***	0.128***	0.128***
	(0.003)	(0.002)	(0.001)	(0.001)
Summer 1	0.024***	0.023***	0.017***	0.018***
	(0.005)	(0.005)	(0.002)	(0.002)
Grade 1	0.082***	0.082***	0.084***	0.084***
	(0.001)	(0.001)	(0.001)	(0.001)
Summer 2	-0.010**	-0.009**	-0.015***	-0.014***
	(0.004)	(0.003)	(0.003)	(0.002)
Grade 2	0.056***	0.057***	0.057***	0.058***
	(0.001)	(0.001)	(0.001)	(0.001)
Difference between private and public schools				
Score, start of kindergarten	0.429***	0.449***	0.391***	0.401***
	(0.054)	(0.052)	(0.046)	(0.045)
Monthly learning rates				
Kindergarten	-0.015**	-0.017***	-0.016***	-0.016***
	(0.005)	(0.004)	(0.002)	(0.003)
Summer 1	0.006	-0.004	0.002	0.001
	(0.012)	(0.009)	(0.008)	(0.008)
Grade 1	-0.007	-0.006	-0.009***	-0.009***
	(0.005)	(0.003)	(0.003)	(0.003)
Summer 2	0.012	0.006	0.012	0.014
	(0.011)	(0.008)	(0.008)	(0.008)
Grade 2	-0.005	-0.003	-0.004	-0.005
	(0.003)	(0.003)	(0.003)	(0.003)
Scores	31,268	36,654	31,268	36,654

*p<0.05, **p<0.01, ***p<0.001. School-clustered standard errors in parentheses. The summer differences between public and private students are highlighted to draw attention.