January 2021

Bridging the Covid Divide

How States Can Measure Student Achievement Growth in the Absence of 2020 Test Scores

By Ishtiaque Fazlul, Cory Koedel, Eric Parsons, and Cheng Qian

Foreword by Michael J. Petrilli

THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

About the Fordham Institute

The Thomas B. Fordham Institute promotes educational excellence for every child in America via quality research, analysis, and commentary, as well as advocacy and exemplary charter school authorizing in Ohio. It is affiliated with the Thomas B. Fordham Foundation, and this publication is a joint project of the Foundation and the Institute. For further information, please visit our website at www.fordhaminstitute.org. The Institute is neither connected with nor sponsored by Fordham University.

Suggested Citation for This Report

Ishtiaque Fazlul, Cory Koedel, Eric Parsons, and Cheng Qian. *Bridging the Covid Divide: How States Can Measure Student Achievement Growth in the Absence of 2020 Test Scores.* Washington D.C.: Thomas B. Fordham Institute (January 2021). https://fordhaminstitute.org/national/research/bridging-covid-divide-how-states-can-measure-student-achievement-growth-absence

Acknowledgments

This report was made possible through the generous support of the Walton Family Foundation and our sister organization, the Thomas B. Fordham Foundation. We are grateful to authors Ishtiaque Fazlul, Cory Koedel, Eric Parsons, and Cheng Qian for their deep expertise, thoroughness, and adherence to aggressive timelines.

External reviewer Kata Mihaly of the RAND Corporation provided helpful and detailed feedback on the draft report. We also want to thank Adam Tyner for managing the project, Chester E. Finn, Jr. for reviewing drafts, Victoria McDougald for overseeing media relations, Olivia Piontek for handling funder communications, Pedro Enamorado for managing report production, and Fordham research assistant Tran Le and intern Trinady Maddock for providing assistance at various stages in the process. Finally, we thank Dave Williams for developing the report's layout and design, and Pamela Tatz for copyediting the report.

Cover Photo: ThamKC/iStock/Getty Images

Contents

Foreword

by Michael J. Petrilli

The Covid-19 pandemic has run roughshod over so much of our education system, closing schools, sending students home to try to learn remotely, and obliterating last year's summative state tests. One consequence of that cancellation is that even if students are tested this spring, it will still be impossible to construct typical measures of their learning growth, as most such measures incorporate the previous year's score. As fanatics for student growth measures—given that they are the fairest and most accurate metrics of schools' impacts on achievement—we wanted to know if some kind of value-added calculations could still be produced despite the testing gap year. Such measures would provide helpful information about which districts, schools, and students have progressed the most and which have experienced the worst Covid-induced learning loss during the pandemic. That would help us identify schools and practices worth emulating, and highlight institutions where students need the most help once Covid-19 is behind us.

To investigate the feasibility of this, we turned to a team of researchers in the department of economics at the University of Missouri—Ishtiaque Fazlul, Cory Koedel, Eric Parsons, and Cheng Qian—who have many years of experience studying how best to measure achievement growth. The team used administrative data from Missouri to simulate the testing gap year that states face as a result of Covid-19, and to generate ideas about how to work through it. Using data from the 2016–17 through 2018–19 school years, they calculated growth over two years to determine how similar gap-year estimates are to the "business-as-usual" condition where testing data are available every year.

Their results, explained in detail in this report, speak to the feasibility of estimating two-year growth measures for districts and for schools, including technical suggestions for handling thorny data issues. The researchers also go on to assess the feasibility of growth measures when *two* years of test scores are missing, simulating the condition if spring 2021 testing is also cancelled.

There's good news and bad news.

- **Happily, both district- and school-level growth estimates based on a single-year gap convey information that's very similar to growth estimates based on data with no gap year.** Rankings of districts and schools only change slightly when a gap year in testing is simulated, and demographic factors such as race and socioeconomic status are not predictive of such changes. This analysis also suggests that such estimates will be valid for large subgroups such as economically disadvantaged students. (We can't say whether that will be the case for smaller subgroups.)
- **But there's bad news, too. Just 27 percent of students attend schools that could generate growth measures if *two* consecutive years of tests are missing.** That's because most students in the standard testing window (grades three through eight) who were tested in 2019—the last time statewide assessments were implemented—will be in different schools by the spring of 2022. (For example, third graders in 2019 will be sixth graders in

2022, which in most districts will make them middle schoolers.) If we want to have any school-level measures of student progress in the near future, it's vitally important that states assess students in 2021. (District growth measures will be doable, and relatively accurate, with another year of missing test data.)

In practical terms, what does this mean?

Calculating student growth measures from 2019 to 2021 is eminently feasible, and the results will be quite accurate—so long as states test students this year. Those measures will provide essential information to guide the educational recovery phase.

But if states cancel testing this year, too, it will be extremely difficult to determine how effective individual schools were during this challenging, historic period in American education. And of course, it will further delay the time until we can restart measuring student progress and holding schools accountable again.

To be sure, we understand the challenge of testing students during a pandemic. Though the miraculous vaccines are offering light at the end of a long and dark tunnel, it's hard to predict exactly how the next few months will unfold. Even if teachers are vaccinated and students are welcomed back for in-person learning, some families will likely want their children to remain home until *they* are vaccinated, too. And taking precious days out of the instructional calendar for testing this spring—just when schools can finally start to address students' social and emotional needs, and significant learning losses—is a hard sell, even for testing-and-accountability hawks like us.

So allow me to make a humble suggestion (albeit one not proposed by the study's authors): States should shift the spring 2021 assessments to fall 2021 when schools reopen. This will allow them to compute those all-important student growth measures for the 2019–21 period, plus establish a baseline for student progress during the 2021–22 school year. To be sure, some states will be better equipped to manage this move than others, particularly those that don't now legislatively mandate testing in the spring and that have enough internal capacity to acclimate schools effectively to new fall assessment schedules.

If they start now, they've got nine months to put revised policies in place. With mere weeks to throw together plans when Covid-19 first descended in March 2020, that should feel like a lifetime to state and local education officials today.

Introduction

State testing programs across the U.S. were halted in the spring of 2020 in response to the Covid-19 pandemic. Virus-induced uncertainty in the education system has persisted into the 2020–21 academic year, with states and districts implementing—and often modifying—a variety of instructional models ranging from fully remote to fully in person, with hybrid options in between. As of this writing, it is still unclear which states will conduct end-of-grade testing this year.[1]

Where testing does resume in spring 2021, the question of whether and how to resume the measurement of achievement growth will arise. Measuring growth is obviously not the top priority for state and local education agencies navigating the pandemic, but it remains important—arguably more so in light of the circumstances. This is because growth data are widely viewed as the most reliable data for assessing individual student learning and the efficacy of districts and schools. Families and educators have responded in a variety of ways to the pandemic, resulting in a diverse range of student learning experiences. Growth data are our best hope for understanding which students have been most affected by the pandemic, where learning deficiencies are most pronounced, and which district and school policy responses have been most effective in minimizing the negative consequences of Covid-induced instructional changes.

However, the 2020 testing gap poses a clear problem for measuring test-score growth. Traditional growth measures from 2018–19 to 2019–20 and 2019–20 to 2020–21 cannot be constructed. Given the testing gap, we consider the prospects for estimating growth over a two-year period, assuming that states resume testing this spring. The "gap-year" growth data would span the period from spring 2019 to spring 2021. Considering the possibility that some states may cancel a second consecutive year of testing, we also conduct a brief analysis to assess the prospects for estimating growth over a three-year period.

> **"Growth data are our best hope for understanding which students have been most affected by the pandemic..."**

Our objective is to gain insight into how well a gap-year growth model using data from 2019 and 2021 will perform compared to a hypothetical "business-as-usual" condition in which annual testing data were available from spring tests in 2019, 2020, and 2021. To do this, we analyze administrative data from Missouri spanning the pre-Covid 2016–17 to 2018–19 school years. We simulate the data condition of a gap year in testing by artificially censoring the 2017–18 test data *as if* a gap year had occurred, and we then produce two-year growth estimates using spring tests from 2017 and 2019 for schools and districts. Because there was no true test gap, we can compare growth estimates based on the gap-year data to growth estimates obtained using all of the annual testing data over the sample period. This allows us to document the consequences of the gap year in terms of deviations of the growth estimates from the business-as-usual condition.

We focus our analysis on the prospects for estimating school- and district-level growth metrics. These metrics are important for several reasons. First, for the many states that incorporate growth into school and district accountability systems, understanding how well we can model

growth with a gap year in the data is important for making decisions about these systems in 2021. Second, variation in school- and district-level growth during the pandemic can be used to identify locales where students have been the most and least impacted by Covid. This is important information for planning intervention and remediation. Moreover, because school- and district-level growth measures can be linked with data on schools' and districts' instructional responses to the pandemic, they can be used to assess the efficacy of different strategies for minimizing the learning costs of Covid.

A caveat to our analysis is that it is most directly informative about a scenario where there is a missing year of test data, but after the gap year, we return to a condition in which (virtually) all students are tested. Currently there is still uncertainty about whether testing will happen in spring 2021, and if it does, there is even more uncertainty about which students will be tested and via what mode (e.g., in person or online). Our view is that there is too much uncertainty at this juncture over these questions for *ex ante* modeling of this process to be fruitful, but this issue bears monitoring as spring 2021 testing rolls out. Regardless, our analysis establishes a baseline degree of the credibility of gap-year growth estimates, pending the assessment of other aspects of state testing efficacy this spring.

Methodology

Estimating value added

We estimate school- and district-level growth using value-added models (VAMs) based on student test scores in math and English language arts (ELA) during grades 4 through 8.[2] We consider two modeling structures: (1) a one-step VAM and (2) a two-step VAM. Both modeling structures are common in research and policy applications.[3] Within each structure, we estimate models with varying degrees of controls for student characteristics and school/district circumstances.

Table 1 describes the models and their features.[4] The sparse models in columns (1) and (2) control for just lagged student achievement in math and ELA. We control for lagged achievement in both subjects, and students must have a lagged score in the same subject to be included in the analytic dataset. The models with student controls in columns (3) and (4) additionally include indicator variables for student race/ethnicity, gender, free/reduced-price lunch status, English language learner status, special education status, and mobility status. For the latter variable, we define a student as "mobile" if the student is observed for less than one year in the school where the test is taken. The model in column (5) additionally includes school- and district-level averages of the lagged test-score and student-characteristic variables to control for the schooling environment.

Table 1. The five value-added models we study differ by the extent of the control variables included and the modeling structure.

Structure	(1)	(2)	(3)	(4)	(5)
	1-Step, Sparse	2-Step, Sparse	1-Step, Student Controls	2-Step, Student Controls	2-Step, All Controls
Student Lagged Test Scores (Math and ELA)	✓	✓	✓	✓	✓
Individual Student Characteristics			✓	✓	✓
School- and District-Average Student Characteristics					✓

Notes. All models also include fixed effects for student grade levels. The individual student characteristic controls are for race/ethnicity, gender, free/reduced-price lunch eligibility status, English language learner status, special education status, and mobility status. The school- and district-average characteristics are of these same variables and lagged achievement, to control for the schooling environment. The precise equations describing these models are available in Fazlul et al. (2021).

We attribute student growth to the contemporary school or district in all models. This is the common approach under normal circumstances—i.e., growth from year (t–1) to year t is attributed to the year t school or district. In the gap-year models, this is a potential concern because there is extra mobility during the gap year. In our baseline estimation condition, we do not make any adjustments to the gap-year models to account for the extra mobility, but later we examine the sensitivity of our findings to such adjustments.

Two aspects of model heterogeneity are noteworthy. First, the models differ in the level of control variables included, ranging from the sparse models in columns (1) and (2) of Table 1 to the very rich specification with detailed student-, school-, and district-level controls in column (5). Second, the one-step and two-step structures differ in how aggressively they control for student-, school-, and district-level factors that may influence test scores. The two-step model is more aggressive in this regard.[5]

All of the value-added estimates from these models are "shrunken" *post hoc* to reduce the influence of estimation error. Shrinkage is a statistical procedure that pulls (i.e., shrinks) the estimates for individual schools and districts toward the average, with the power of the pull dependent on the reliability of the individual estimate. The end result is that small schools and districts will be pulled toward the mean more strongly than their larger counterparts. This embodies the Bayesian notion that in the absence of information to the contrary, our best guess is that any individual school or district is "average" in terms of value added. Shrinkage is a common tool used in the value-added literature.[6]

It is beyond the scope of this report to delve into the technical and policy tradeoffs of the various models, but interested readers can find discussions of these tradeoffs in the research literature on value-added modeling.[7]

Simulating the gap year in testing

For each of the models shown in Table 1, we estimate value added with and without the gap-year data censoring in place. We begin by using the uncensored data to estimate two consecutive value-added estimates for each unit (either a school or a district) with data from 2016–17 to 2017–18 and 2017–18 to 2018–19. We then sum the two single-year estimates to produce an estimate of value added over the two-year period. This process is meant to replicate how a typical system would estimate value added over two years, assuming all data were available.

Next, to simulate the gap year in testing, we censor the 2017–18 test data and directly estimate value added using data from 2016–17 and 2018–19. This scenario is meant to reflect the data condition assuming that testing resumes in spring 2021—i.e., the condition of a gap year between the 2018–19 and 2020–21 tests. By comparing the full-data scenario to the gap-year scenario, we assess the potential for the gap-year models to recover accurate estimates of test-score growth over the two-year period.

We also compare our gap-year estimates using the censored data to value-added estimated for the single year from 2017–18 to 2018–19. Mapping this scenario to the current situation with Covid, it informs the following question: How effective would it be if gap-year growth from 2018–19 to 2020–21 were used to proxy for what growth would have been from 2019–20 to 2020–21? We do not view this as a desirable policy use of the censored data, but we show the results in the appendix for completeness. (*Appendix A* also describes why we are not in favor of this approach.)

Finally, we extend our analysis to simulate the effect of two consecutive gap years in testing, which will happen if testing does not occur in spring 2021 and resumes in spring 2022. For this extension, we censor the test data in our panel in 2016–17 and 2017–18 and calculate growth from 2015–16 to 2018–19. We then compare growth estimated over the three-year period to the analogous full-data condition, where three-year growth is calculated as the sum of annual growth estimates from 2015–16 to 2016–17, 2016–17 to 2017–18, and 2017–18 to 2018–19.

Data

Our analysis is based on administrative microdata from Missouri covering all students tested in grades 3 to 8 in math and ELA during the school years 2015–16 through 2018–19. Hereafter, we identify schools by the spring year (e.g., 2018–19 as 2019).[8]

We standardize student test scores throughout by grade-subject-year and estimate growth for all districts and schools with at least ten students in each model and scenario. When we correlate and otherwise compare growth estimates using the full data and gap-year data, the comparisons are restricted to districts and schools that meet the size threshold in both data conditions. Only very small Missouri districts and schools are omitted from our analysis due to the sample size restriction.[9]

We do not expect contextual features of Missouri to limit the generalizability of our findings in most respects. That said, two aspects of the Missouri data merit mention. First, Missouri changed its math and ELA tests once each between 2016 and 2019. A previous study explored the impact of test-regime changes on value-added estimates in math and ELA across multiple states and found that such changes typically do not affect model performance substantively.[10] Moreover, we have performed internal diagnostic work using the Missouri data specifically that supports this inference.[11]

Second, Missouri is a "small district" state, and growth estimates for smaller districts will be more sensitive to data changes because they have fewer students to balance out the sampling variance that the data changes create. To improve the generalizability of our findings to states with larger school districts, we also produce separate estimates for a subsample of the 100 largest districts in Missouri. Table 2 summarizes our data in terms of students, schools, and districts.

Table 2. Summary statistics for students, schools, and districts in the analytic sample

	Mean	SD	Min	Max
Student information				
Standardized Math Score	0.016	0.991	−5.978	5.281
Standardized ELA Score	0.016	0.991	−5.389	6.146
Asian	0.021	0.144	0.000	1.000
Black	0.156	0.363	0.000	1.000
Hispanic	0.065	0.247	0.000	1.000
White	0.714	0.452	0.000	1.000
Multiple and Other Race/Ethnicity	0.044	0.197	0.000	1.000
Female	0.490	0.500	0.000	1.000
Eligible for Free/Reduced-Price Lunch	0.521	0.500	0.000	1.000
English Language Learner	0.049	0.217	0.000	1.000
Individualized Education Program	0.130	0.336	0.000	1.000
Mobile Student	0.040	0.196	0.000	1.000
School Information				
Urban	0.175	0.380	0.000	1.000
Suburban	0.239	0.427	0.000	1.000
Rural/Town	0.496	0.500	0.000	1.000
Enrollment	357.0	217.4	12.0	1,728.0
District Information				
Enrollment (All)	1,603.2	3,194.1	18.0	24,955.0
Avg. Number of Schools (All)	4.2	6.1	1.0	76.0
Enrollment (Large District Subsample)	6,321.8	5,332.5	839.0	24,955.0
Avg. Number of Schools (Large District Subsample)	12.4	11.0	2.0	76.0
N (Student Years, 2017–19)	972,877			
N (Unique Schools, 2017–19)	1,730			
N (Unique Districts, 2017–19)	557			

Notes. These summary statistics are based on the analytic sample of students in grades 4 to 8 with lagged test scores in 2016–17, 2017–18, and 2018–19 who attend districts and schools with at least ten test takers. Urbanicity information is taken from the 2018–19 Common Core of Data. The large-district subsample is selected to include the 100 districts in Missouri with the largest populations of test takers included in the gap-year model. Other size-based selection criteria produce a similar sample; we chose this criterion in order to isolate districts in Missouri with the largest samples relevant for our primary analysis.

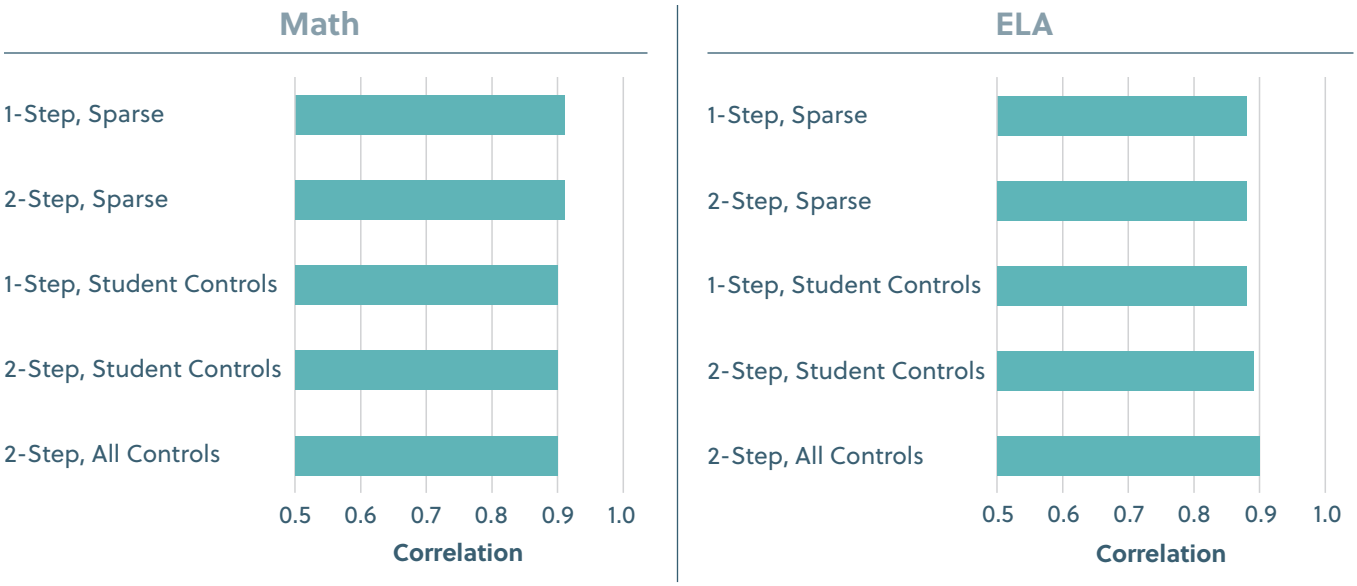
Results

What happens to district-level growth data with a gap year?

Finding 1: **District growth estimates based on a single-year gap convey similar information to growth estimates based on data with no gap year.**

We begin in Figure 1 by reporting correlations between district-level growth estimates over the two-year period spanning the gap year, with and without the gap-year data censoring. The growth models correspond to those described in Table 1 (above). Each bar represents the correlation between the output of the specified model with and without data censoring in place. Recall that in the full-data condition, we estimate value added from 2017 to 2018 and then 2018 to 2019 and sum the two, to arrive at an estimate of two-year growth. With the gap year, we directly estimate growth from 2017 to 2019.

Figure 1. District-level growth estimates over the two-year period spanning the gap year, with and without the gap-year data censoring, are highly correlated.



Note. For details on the models, see *Methodology*.

Figure 1 shows that the growth estimates are highly correlated—the correlation coefficients across models in both math and ELA are consistently around 0.90. The correlations for the math models are slightly higher. A summary takeaway is that growth estimates based on gap-year data convey similar information to growth estimates based on all the data.

The correlations in Figure 1 are high, but one might wonder why they aren't even higher. After all, both the full-data and gap-year estimates are meant to capture growth over the same two-year period. There are two primary reasons that the growth estimates differ. First, the cohorts of students who contribute to the estimates in each condition are not identical. For example, students in the eighth grade in 2018 (the censored year) will contribute to business-as-usual growth estimated with the full dataset from 2017 to 2018 but will not contribute to gap-year growth because in 2019 they are outside of the tested grade span. Second, there is some gap-induced modeling and estimation variance—most notably, the predictive power of lagged achievement over contemporary achievement varies depending on the presence of the gap year, and correspondingly, this can affect the predictive power of the other control variables.[12]

Finding 2: The overwhelming majority (86 to 88 percent) of districts do not change ranking categories due to the gap year in the data.

Table 3 uses transition matrices to document the stability of district rankings based on test-score growth in math (panel A) and ELA (panel B) between the business-as-usual and gap-year conditions. The transition matrices emphasize ranking changes in the tails of the distribution—i.e., the top and bottom 10 percent of districts—with the rationale that many policy applications focus on the distribution tails rather than the center. The rows of each matrix indicate district placements in the full-data growth rankings, divided into groups of "bottom 10 percent," "middle 80 percent," and "top 10 percent." The columns divide districts into the same groups based on their rankings using gap-year growth estimates.

Table 3. Transition matrices show that the ranking category does not change for most districts due to the gap year in testing (results shown only for the two-step all-controls model for brevity).

Panel A. Math

		Gap-year Data Growth Ranking		
		Bottom 10 Percent	Middle 80 Percent	Top 10 Percent
Full Data Growth Ranking	Bottom 10 Percent	6.7	3.3	0.0
	Middle 80 Percent	3.3	73.9	2.8
	Top 10 Percent	0.0	2.8	7.2

Panel B. ELA

		Gap-year Data Growth Ranking		
		Bottom 10 Percent	Middle 80 Percent	Top 10 Percent
Full Data Growth Ranking	Bottom 10 Percent	6.3	3.7	0.0
	Middle 80 Percent	3.7	73.0	3.3
	Top 10 Percent	0.0	3.3	6.7

Notes. Each cell indicates the percentage of Missouri districts for which the ranking profile matches the row and column headers. The transition matrices for the other VAMs are very similar to the transition matrices shown here. They are suppressed for brevity but can be found in Fazlul et al. (2021).

The value in each cell indicates the percent of districts for which the rankings in the full-data and gap-year conditions fit the profile indicated by the row and column. The sum of the cells in each transition matrix is 100 by construction. If the full-data and gap-year models produced identical output, the diagonal elements of the transition matrices would have values of 10-80-10 and the off-diagonal elements would have values of zero. Thus, nonzero off-diagonal entries provide an indication of disagreement between the growth estimates. In this way, the transition matrices are complementary to the correlations presented above.[13]

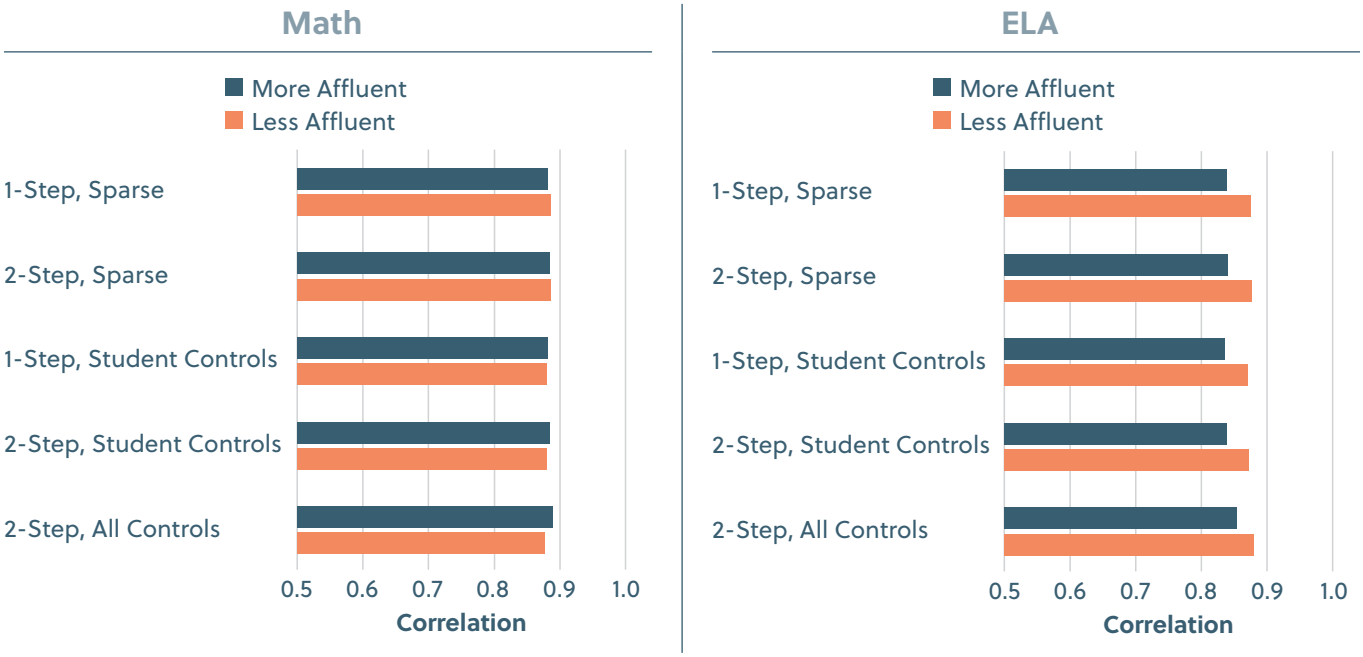
The overwhelming majority of the weight in each transition matrix in Table 3 is on the diagonal—86 to 88 percent of districts remain in the same ranking category regardless of whether the full data or gap-year data are used. The districts that change categories are relatively close to the 90th and 10th percentile cutoffs; among these districts, the average value of the percentile ranking change is about ten percentile points—e.g., a district that moves from the 85th to 95th percentile in the ranking distribution between models.

Finding 3: Student subgroup growth estimates based on gap-year data can be reliably estimated.

To get a sense of the applicability of these methods for estimating district-level growth by student subgroup, we split the sample by student socioeconomic status and conduct the same analysis as above (Figure 2).[14] The results show the gap-year and full-data growth estimates are highly correlated in the split samples. The correlations are attenuated relative to Figure 1, but this is likely the result of the fact that we're using smaller samples in the split analysis than in the full analysis.

We conclude from these results that differences in student growth across student subgroups can be tracked using gap-year growth data with only modest consequences in terms of estimation accuracy. Note that state-aggregated subgroup comparisons will be even easier to make in order to assess statewide growth gaps along various dimensions.

Figure 2. For student data disaggregated by socioeconomic status, the correlations remain high for the district-level growth estimates with and without the gap-year data censoring.



Note. We refer to students coded as ineligible for free or reduced-price lunch as "more affluent" and those who are coded as eligible as "less affluent." For details on the models, see *Methodology*.

Finding 4: Most of the changes to districts' growth rankings caused by the gap year are not associated with observable district characteristics.

Next, we assess whether district characteristics systematically predict ranking changes between the full-data and gap-year conditions. There is not a strong theoretical reason to expect districts' characteristics to predict the direction of growth ranking changes due to the gap year. For example, as noted above, the changes are partly driven by the use of nonoverlapping cohorts to produce the full-data and gap-year estimates. Unless one believes that some cohorts of students (e.g., third-grade students in 2018) are systematically different from other cohorts in the same districts (e.g., fourth-grade students in 2018), this source of ranking discrepancies should not be consistently predictable with district characteristics. As for the other source of ranking discrepancies—modeling and estimation variance—*ex ante* we do not expect district characteristics to systematically drive fluctuations via this channel either. However, *ex post*, after the model is estimated, changes in the coefficients can lead to ranking changes that are correlated with district characteristics.

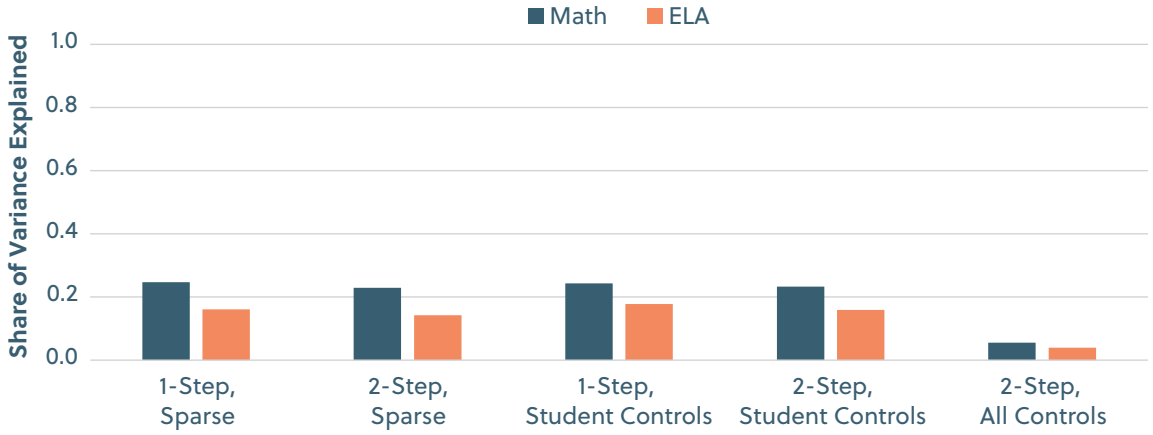
To test whether district characteristics predict ranking changes, we regress the ranking percentile change—i.e., the growth ranking percentile based on the gap-year data minus the growth ranking percentile based on the full data—on the following district characteristics: the district average same-subject test score in 2017, racial-ethnic composition, district enrollment, and the shares of

students who (a) are eligible for free or reduced-price lunch, (b) are classified as English learners, (c) have an individualized education program (IEP), and (d) are mobile. We use two definitions for mobility. The first is our standard measure, which is whether the student attended the school where the year-*t* test occurred for less than one year. The second is the share of students who attended more than one district in 2018 and 2019 (i.e., the gap year and subsequent year). These students are of particular interest for the comparison of models because the gap-year models, under the baseline estimation conditions without any adjustments for student mobility, will have more misclassification errors because they attribute all growth over the two-year period to the year-*t* district.

We also regress the absolute value of the growth ranking changes on the same district characteristics to look for predictors of ranking volatility. There is also little reason to expect most district characteristics to predict ranking volatility between models, with the exceptions of (a) district enrollment, because larger districts should be less sensitive to sample changes caused by the gap year, and (b) the share of students who attended more than one district in 2018 and 2019, because the models differ in how they attribute growth to districts for these students.

Figure 3 shows the R-squared values from our regressions of the directional ranking changes on district characteristics, which indicate the fraction of the variance in ranking changes that can be explained by our host of district characteristics. In the first four models in the figure (one-step sparse through two-step student controls), the R-squared values are in the range of 0.23 to 0.25 in math and 0.14 to 0.18 in ELA. This means that about 75 to 78 percent of the ranking changes in the math models and 82 to 86 percent of the ranking changes in ELA are not explained by any of the observable district characteristics. Even less of the variance in ranking changes—only about 5 percent in math and 4 percent in ELA—is explained by district characteristics in the two-step all-controls model, highlighting an advantage of this approach. The finding that most of the ranking fluctuations are unexplained by observable district characteristics—in all models, but especially the two-step all-controls model—is consistent with our theoretical prediction stated above.[15]

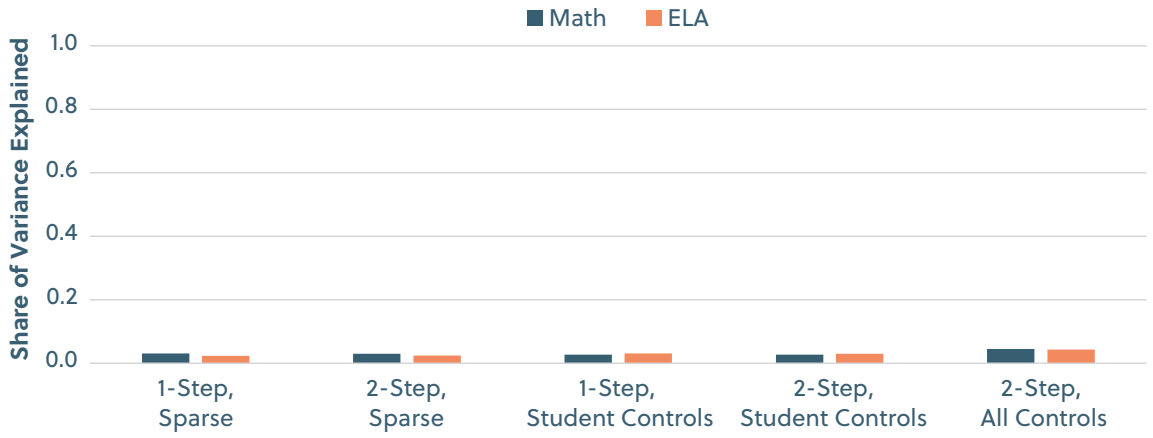
Figure 3. Most of the changes to districts' growth rankings caused by the gap year are not explained by observable district characteristics. This is especially true for the two-step all-controls model.



Notes. A value of 1.0 on the vertical axis would indicate that observable district characteristics can explain all of the variance in rank changes; a value of 0.0 would indicate that observable district characteristics can explain none of the variance in rank changes. For details on the models, see *Methodology*.

Figure 4 reports corresponding R-squared results from the same regressions, except where the dependent variable is the absolute value of each district's ranking change. Whereas Figure 3 is informative about the directional predictive power of the district characteristics, Figure 4 is informative about their ability to predict ranking volatility (without regard to direction). Figure 4 shows that district characteristics explain almost none of the variance in the volatility of district rankings between models.

Figure 4. Most of the volatility in districts' growth rankings (i.e., the absolute value of the ranking changes) caused by the gap year is not explained by observable district characteristics.

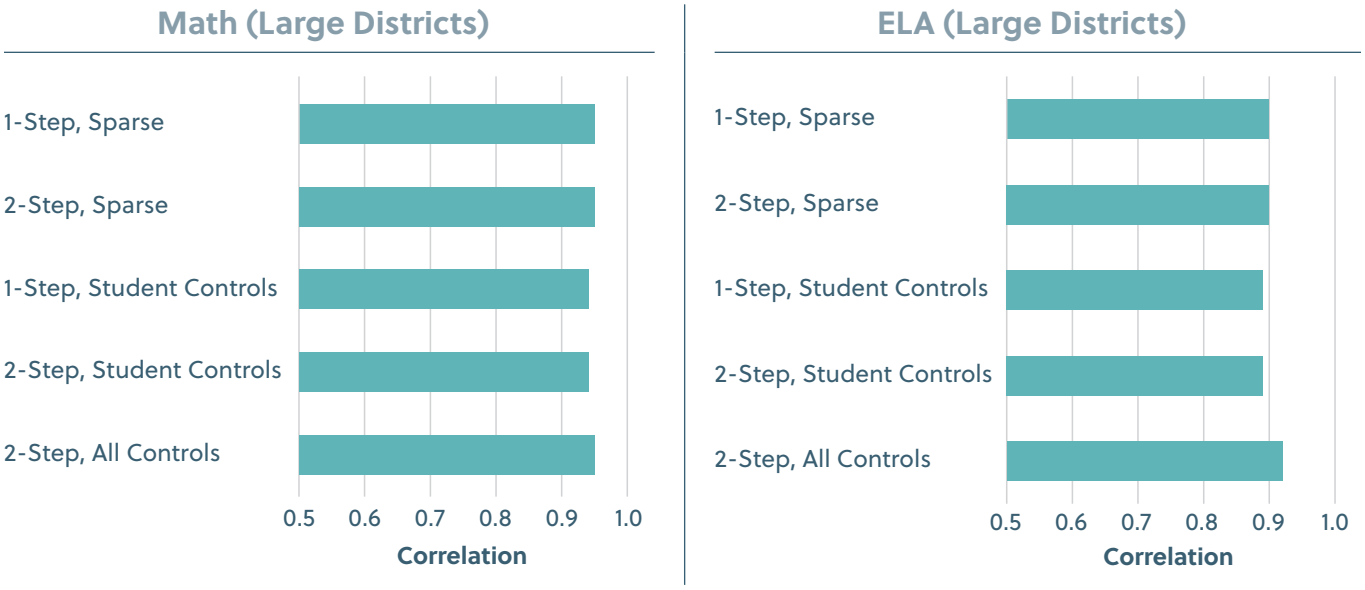


Notes. A value of 1.0 on the vertical axis would indicate that observable district characteristics can explain all of the variance in districts' ranking volatility; a value of 0.0 would indicate that observable district characteristics can explain none of the variance in volatility. For details on the models, see *Methodology*.

Finding 5: The results for the subsample of large districts largely mirror the results for all districts, although the large-district growth estimates using the gap-year data are even more accurate.

Figure 5 replicates the results in Figure 1 for the large-district sample only. The correlations are generally higher for the large districts across models and estimation scenarios. This is expected, given the larger district-level samples. These results show that growth estimates for larger districts will be affected even less by the use of gap-year data relative to the business-as-usual condition.[16]

Figure 5. For the subsample of large districts, the correlations of the district-level growth estimates with and without the gap-year data censoring are even higher.



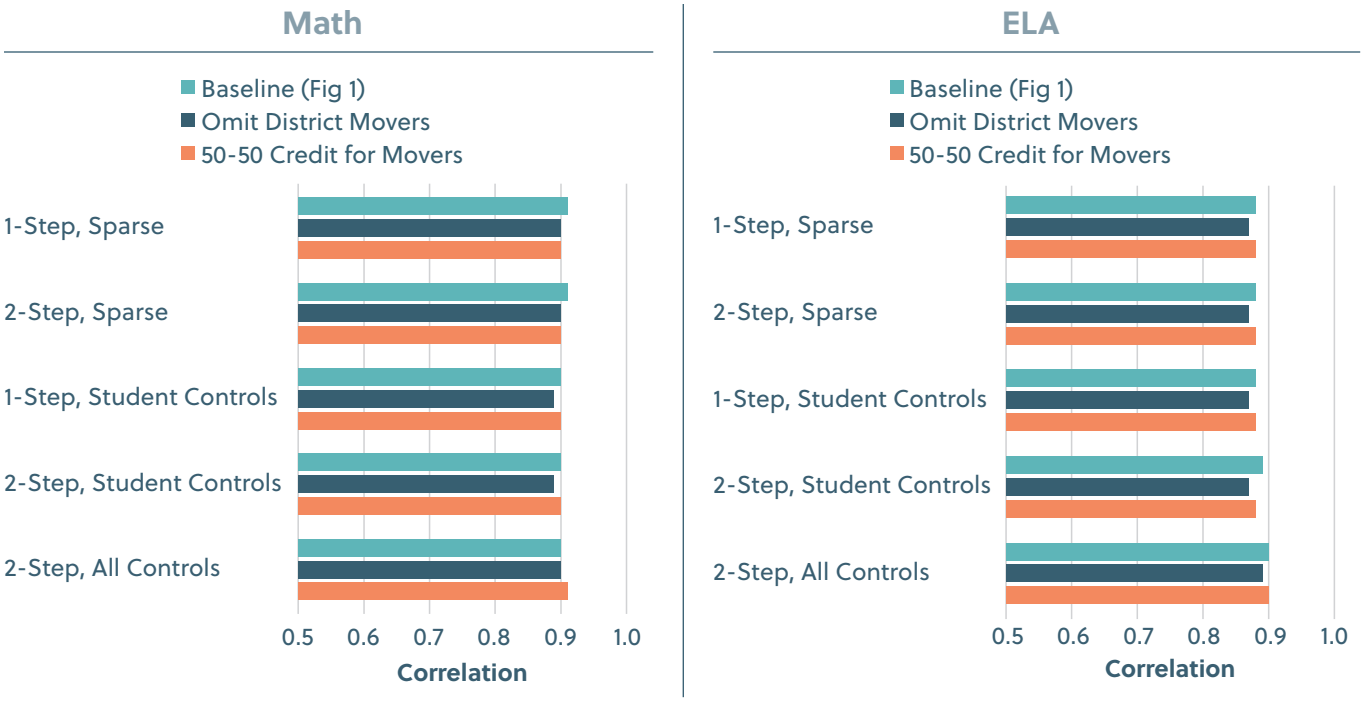
Note. For details on the models, see *Methodology*.

Finding 6: Modifications to the gap-year model to account for student mobility during the gap period do not meaningfully affect growth estimates for districts.

We anticipated that differences in how student mobility is treated might cause our results to differ across data conditions, but we find little evidence of such in our district-level analysis. To illustrate, consider a student who attends District A in 2017 and 2018 but district B in 2019. In the business-as-usual VAM, her growth from 2017 to 2018 will be attributed to District A and her growth from 2018 to 2019 will be attributed to District B. However, using the convention of assigning growth to the contemporary district, in the gap-year model her growth over the full two-year period will be attributed to District B.

We explore the mobility issue empirically in Figure 6, where we compare our baseline correlations from Figure 1 to analogous correlations after we make adjustments to the gap-year models for student mobility. We consider two data adjustments, which are both applied to the gap-year models only. In the first, we drop all students who were not enrolled in the same district in period *t*–1 and *t*—i.e., in 2018 and 2019 in our censored dataset. These students only attended the contemporary district for one of the two years over which gap-year growth is estimated, meaning that their growth over the full period is partially misattributed. The full-data models still include all students and assign single-year growth to the contemporary district, which would be the business-as-usual approach and is facilitated by the availability of the 2018 test score. In the second mobility modification, we retain all mobile students in the gap-year dataset but assign 50 percent weight to the districts attended in 2018 and 2019, respectively.[17]

Figure 6. Modifications to the gap-year models to account for student mobility have a negligible impact on the accuracy of district-level growth estimates.



Note. For details on the models, see *Methodology*.

Figure 6 shows that although extra mobility over the two-year period is conceptually concerning, corrections to account for mobility differently in the gap-year models do not meaningfully change the accuracy of the district growth estimates. To preview the school-level results—and noting that the cross-school mobility rate is much higher—the shared-credit solution (i.e., 50 percent weight) improves the correlations modestly.

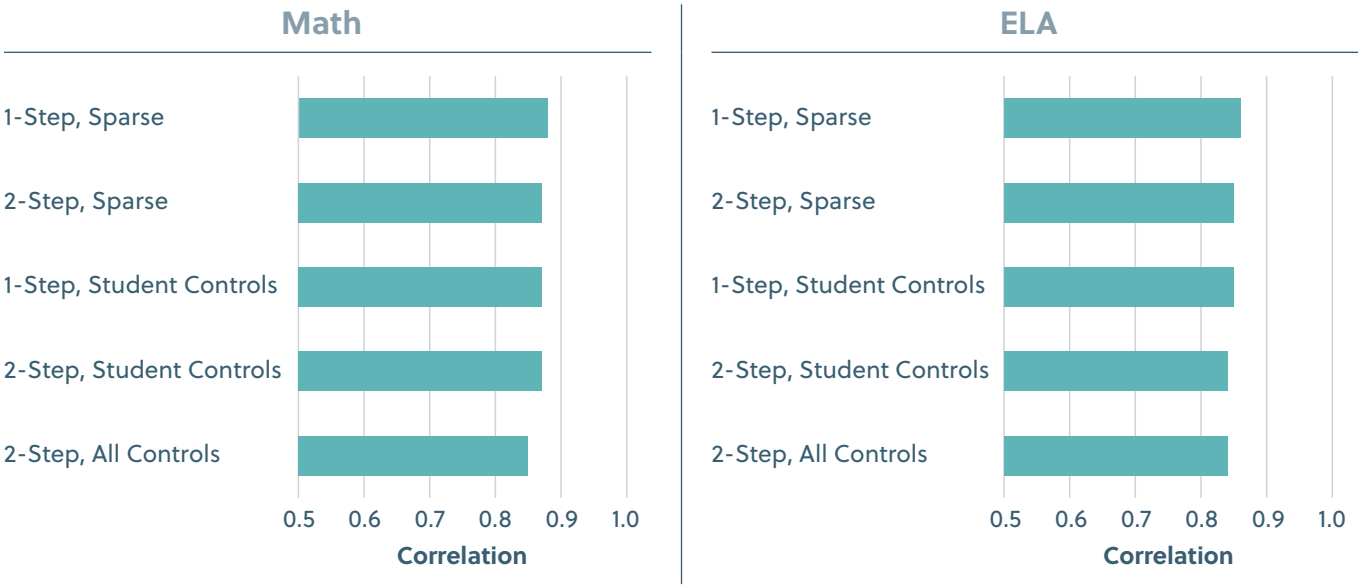
What happens to school-level growth data with a gap year?

Next, we replicate the analysis above for schools. The structure follows from above, except that we focus on our ability to recover school-level growth estimates from the gap-year models.

Finding 7: **School-level growth estimates based on gap-year data convey similar information to growth estimates based on all of the data, although the accuracy of the gap-year estimates is slightly lower for schools than for districts.**

First, Figure 7 replicates for schools what we show in Figure 1 for districts. Figure 7 reports the baseline correlations between school-level growth estimates with and without the gap year, which range from 0.85 to 0.88 in math and 0.84 to 0.86 in ELA, depending on the model. These numbers are lower than the analogous values for districts reported in Figure 1, but they are substantively similar.

Figure 7. School-level growth estimates over the two-year period spanning the gap year, with and without the gap-year data censoring, are highly correlated, although the correlations are lower than for districts.



Note. For details on the models, see *Methodology*.

We also conduct school-level analysis of students by socioeconomic status and find that, as with the district analysis (Figure 2), estimates for more and less affluent students based on gap-year data are highly correlated with estimates based on the full data. (For more on these results, see Figure A3 in *Appendix A*.)

Finding 8: The overwhelming majority of schools (85 to 86 percent) do not change ranking categories due to the gap year in test data.

Table 4 shows transition matrices for schools that correspond to the analogous transition matrices for districts in Table 3. The results are similar—85 to 86 percent of the weight is along the diagonal. This means that the vast majority of schools remain in the same ranking category regardless of whether the full data or gap-year data are used.

Table 4. Transition matrices show that the ranking category does not change for most schools due to the gap year in testing (results shown for only the two-step, all-controls model for brevity).

Panel A. Math

		Gap-year Data Growth Ranking		
		Bottom 10 Percent	Middle 80 Percent	Top 10 Percent
Full Data Growth Ranking	Bottom 10 Percent	6.0	3.9	0.0
	Middle 80 Percent	3.9	72.8	3.3
	Top 10 Percent	0.0	3.3	6.7

Panel B. ELA

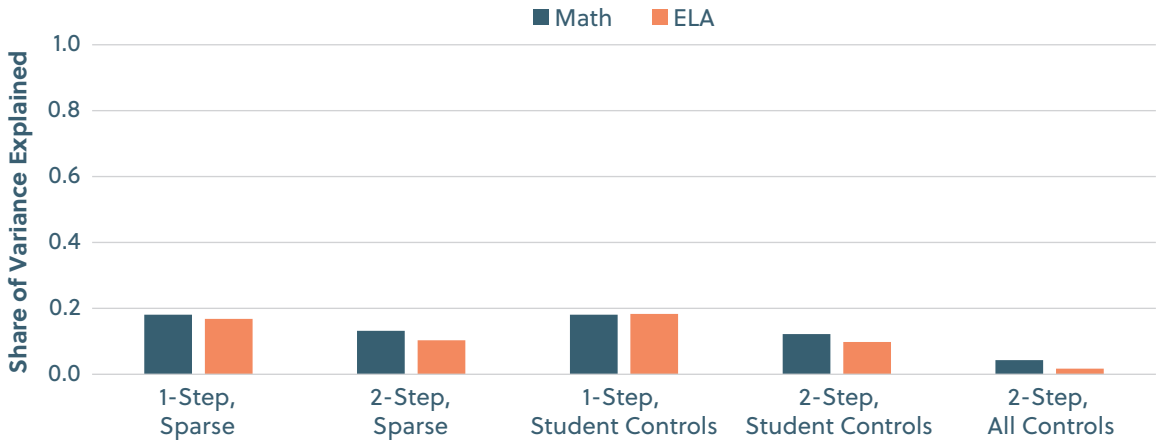
		Gap-year Data Growth Ranking		
		Bottom 10 Percent	Middle 80 Percent	Top 10 Percent
Full Data Growth Ranking	Bottom 10 Percent	6.6	3.3	0.0
	Middle 80 Percent	3.3	73.0	3.7
	Top 10 Percent	0.0	3.7	6.4

Notes. Each cell indicates the percentage of Missouri schools for which the ranking profile matches the row and column headers. The transition matrices for the other VAMs are very similar to the transition matrices shown here. They are suppressed for brevity but can be found in Fazlul et al. (2021).

Finding 9: Most of the changes to schools' growth rankings caused by the gap year are not explained by observable school characteristics.

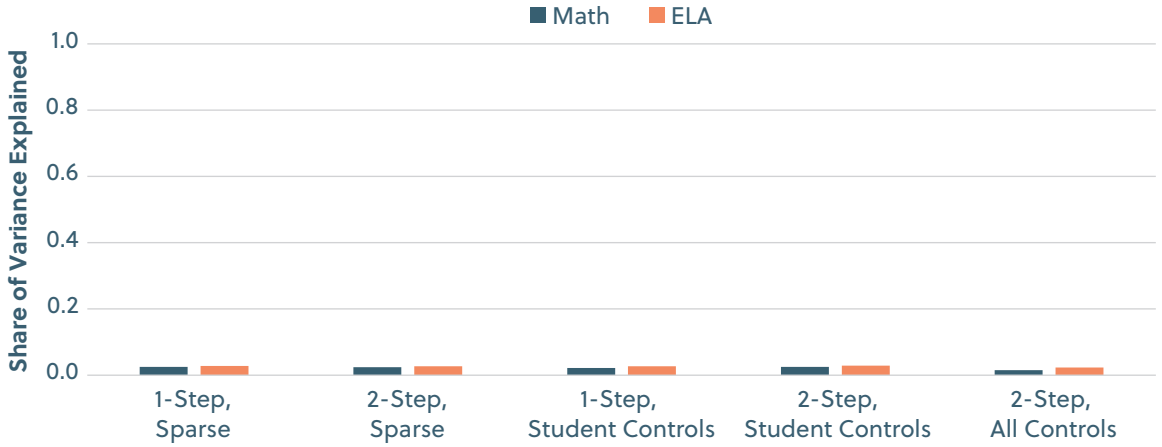
Figures 8 and 9 explore the ability of observable school characteristics to predict changes in school rankings caused by the use of the gap-year data. We use the same regression framework and same observable characteristics as in the district-level analysis (Figures 3 and 4), except that the analysis is conducted at the school rather than district level. Observable school characteristics explain even less of the directional changes in rankings than in the case of districts, but otherwise the patterns of estimates are similar to what we show above for districts.

Figure 8. The variance in rankings as a result of omitting the gap year is not well explained by observable school characteristics, and this is especially true for the two-step, all-controls model.



Notes. A value of 1.0 on the vertical axis would indicate that observable school characteristics can explain all of the variance in rank changes; a value of 0.0 would indicate that observable school characteristics can explain none of the variance in rank changes. For details on the models, see *Methodology*.

Figure 9. Most of the volatility in schools' growth rankings (i.e., the absolute value of the ranking changes) caused by the gap year is not explained by observable school characteristics.



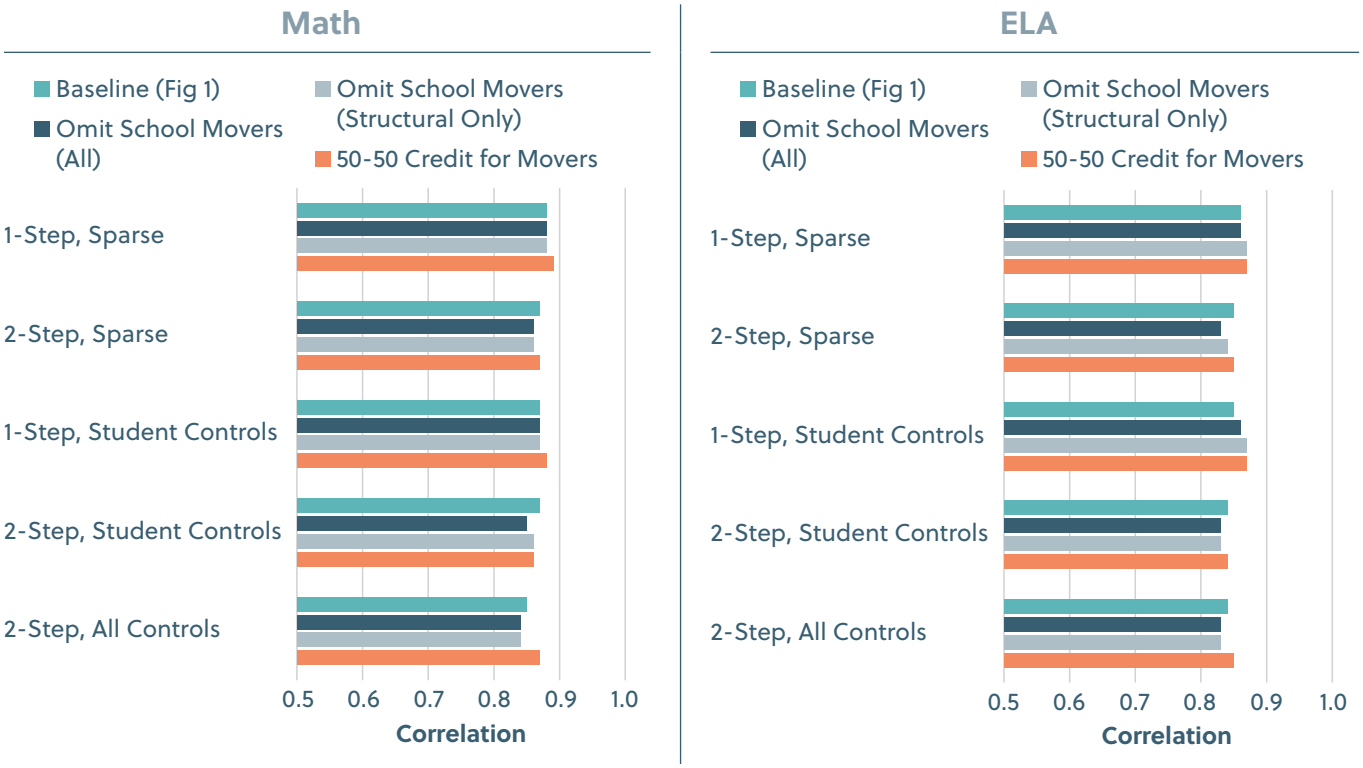
Notes. A value of 1.0 on the vertical axis would indicate that observable school characteristics can explain all of the variance in districts' ranking volatility; a value of 0.0 would indicate that observable school characteristics can explain none of the variance in volatility. For details on the models, see *Methodology*.

Finding 10: For schools, the partial-credit solution to address student mobility during the gap period, where we give 50 percent weight to the schools attended during and in the year after the gap year, modestly improves estimation accuracy relative to the baseline approach.

Figure 10 examines student mobility and replicates the district analysis presented in Figure 6. One addition to the analysis is that we distinguish between structural and nonstructural school moves. A structural school move is a move that occurs because a school's grade span has ended. We define a structural move as occurring when a student is in the terminal grade of his or her school during the gap year—2018 in our simulated environment. A nonstructural move is a move that occurs in a nonterminal grade. A large percentage (about 70 percent) of the school movers in our analytic sample are structural movers.[18]

The results in Figure 10 show that the partial-credit solution, where we give 50 percent weight to the 2018 and 2019 schools, generally improves estimation accuracy relative to the baseline approach. This solution may be more effective in the analysis of schools because there are many more school movers, which makes more accurate attribution of their growth more impactful.

Figure 10. Modifying the gap-year model so that it assigns 50-50 credit to mobile students during the gap year generally improves the accuracy of the growth estimates for schools, albeit modestly.



Note. For details on the models, see *Methodology*.

What happens if there is a second missing year of data?

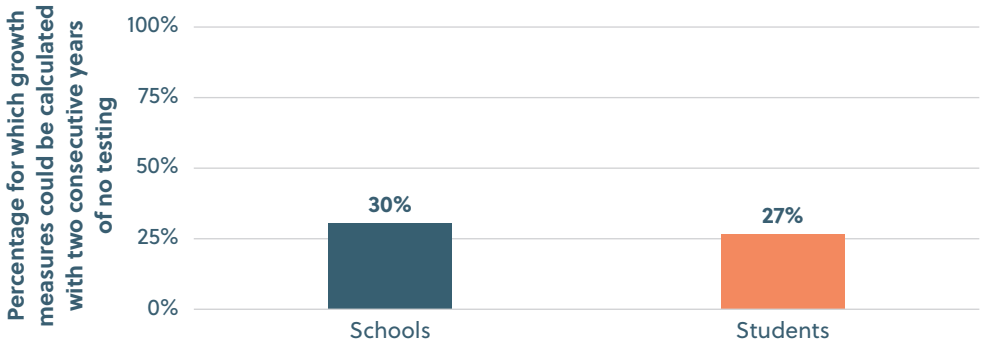
Next, we consider the prospects for estimating test-score growth if state testing is cancelled again in spring 2021. If this were to happen, there would be a two-year data gap, which would necessitate trying to measure growth from 2019 to 2022. In our data, we simulate this situation by further censoring our data panel to remove the 2017 test—i.e., the test gap is expanded to include both 2017 and 2018. We then estimate test-score growth from 2016 to 2019, replicating the "two-year gap" condition.

Finding 11: **The gap-year model when omitting two years of data produces estimates for districts that are highly correlated with estimates based on all of the district data, but the estimates are less accurate than in the case of a one-year test gap.**

Before presenting results for districts, we first discuss a problem with estimating school-level growth in the two-year gap scenario. With a two-year gap, many schools will not have any cohorts of students who remain in the school and are tested spanning the full gap period. We use the 2018–19 Common Core of Data to show this problem by identifying the fraction of U.S. schools for which growth could be estimated with a two-year gap for at least some student cohorts with start- and end-tests in the same school. This requires that the school offer four consecutive grades in the grades 3 to 8 span (e.g., a school offering grades K through 5 would not qualify, as it has only three consecutive grades in this span, while a school offering grades K through 6 would qualify).[19]

Just 30 percent of schools offering any grade in the 3 to 8 range have four consecutive grades in the 3 to 8 span (Figure 11). Further, the schools that meet the criterion of four consecutive grades are smaller on average, and just 27 percent of U.S. students attend such schools. This shows that for most schools, and for the schools attended by most students, we cannot credibly estimate test-score growth with a two-year test gap.

Figure 11. Just 27 percent of students attend schools that could generate growth measures if two consecutive years of tests are missing.

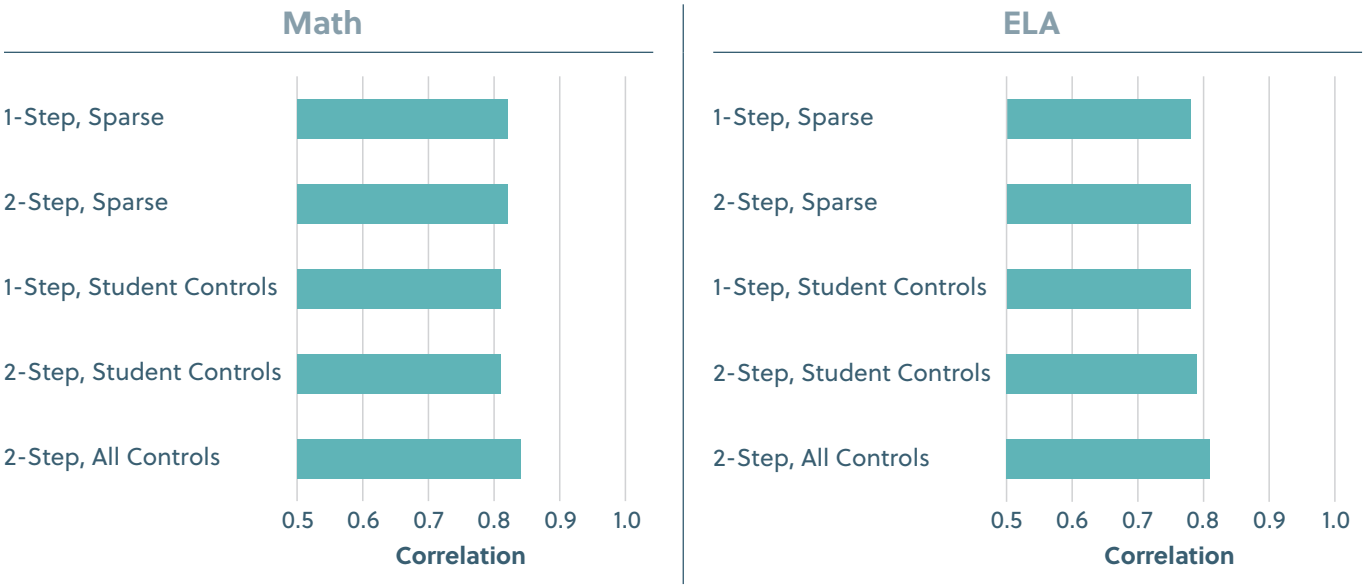


Note. Data are from the U.S. Department of Education Common Core of Data, 2018–19.

Although a school-level growth analysis is infeasible with a two-year gap, a district-level analysis is possible because most districts contain four consecutive grades in the 3 to 8 range. That said, the biggest challenge is that growth can be estimated for even fewer cohorts with a two-year test gap. Specifically, students in grades 3, 4, and 5 in the pre-gap year are the only cohorts for which an endpoint test score would be available to estimate growth (those in grades 6, 7, or 8 in the spring of 2019 will be in high school by 2022 and thus beyond the federal grade-level testing range). Given that lack of cohort overlap is a key driver of discrepancies in district growth estimates with and without the gap year, a prediction is that with a two-year test gap, the discrepancy will be larger.

Figure 12 shows that this is indeed the case. With a two-year test gap, the growth estimates from 2016 to 2019 are less correlated with the three years of summed, single-year estimates than in the case of a one-year test-score gap. The correlations in Figure 12 range from 0.78 to 0.84 for districts, compared to 0.88 to 0.91 in the scenario of a one-year test gap (Figure 1). The correlations are still large and positive, but they indicate a larger degradation of information relative to the full-data condition.

Figure 12. In the case of a two-year test gap, district-level growth estimates over the three-year period spanning the gap years, with and without the gap-year data censoring, are strongly correlated. That said, the gap-year estimates are less accurate than in the case of a one-year test gap.



Note. For details on the models, see *Methodology*.

Implications

We show that growth measures spanning a single year with a gap in data can be estimated and are highly correlated with those estimated based on the full data. This implies that if testing is cancelled for only one year due to Covid (i.e., spring 2020), we have the potential to credibly estimate growth over the two-year period for districts and schools alike, spanning the gap year with testing data from spring 2019 and spring 2021.

While state and local education agencies may decide to pause the use of growth data for accountability purposes during this time, such data would still be useful for understanding where learning deficiencies are most pronounced, which district and school policy responses have been most effective in minimizing the negative consequences of the pandemic, and which students may benefit the most from additional resources and interventions. Importantly, our separate analysis for students by socioeconomic status suggests that districts and schools would still be able to track achievement gaps and monitor equity for larger student subgroups.

Though our research design using a simulated gap year offers a credible approach for predicting the impacts of the 2020 test stoppage on 2021 growth estimates, we return to the caveat mentioned in the introduction that the pandemic, along with schools' and families' varied responses to it, has likely affected which students will be tested in public schools when testing resumes. The compositional effect may also go beyond who is tested and impact the mode through which different students are tested—e.g., within a state, some students may take online tests while others take in-person tests (and whether online and in-person test data can be combined to track student progress is an open question). Due to the high level of uncertainty surrounding the composition effect, we do not attempt to model changes to the composition of test takers caused by Covid *ex ante*. We recommend that when spring 2021 tests are administered, states keep a close eye on test coverage on a district-by-district and school-by-school basis. Depending on the amount of missing data and the degree of selectivity into missingness, test coverage may or may not be a significant problem for producing useful estimates of test-score growth once testing resumes.

Finally, if testing in spring 2021 is also cancelled, our assessment is that it will not be possible to construct useful school-level growth measures spanning the ensuing multiyear gap period, although reasonably accurate district-level growth estimates could still be estimated. In order to estimate school-level growth, testing this year is necessary. If spring testing is still not possible in 2021, some states may consider administering assessments when schools reopen in fall 2021, although such a scenario would pose its own logistical, analytic, and political challenges.

***

The Covid pandemic has continually pummeled our schools with challenges over the past year. Without data from summative assessments this year, we will lose valuable insight into how learning has been affected and which students and schools have been hit hardest. To help students recover learning loss and begin to build our schools back as we move beyond this crisis, good information about student learning must be a top priority.

Appendix A: Additional analysis

In this appendix, we present results from two additional analyses that inform the prospects for using two-year growth as a proxy for a single year of growth, as well as the reliability of student-subgroup growth estimates within schools.

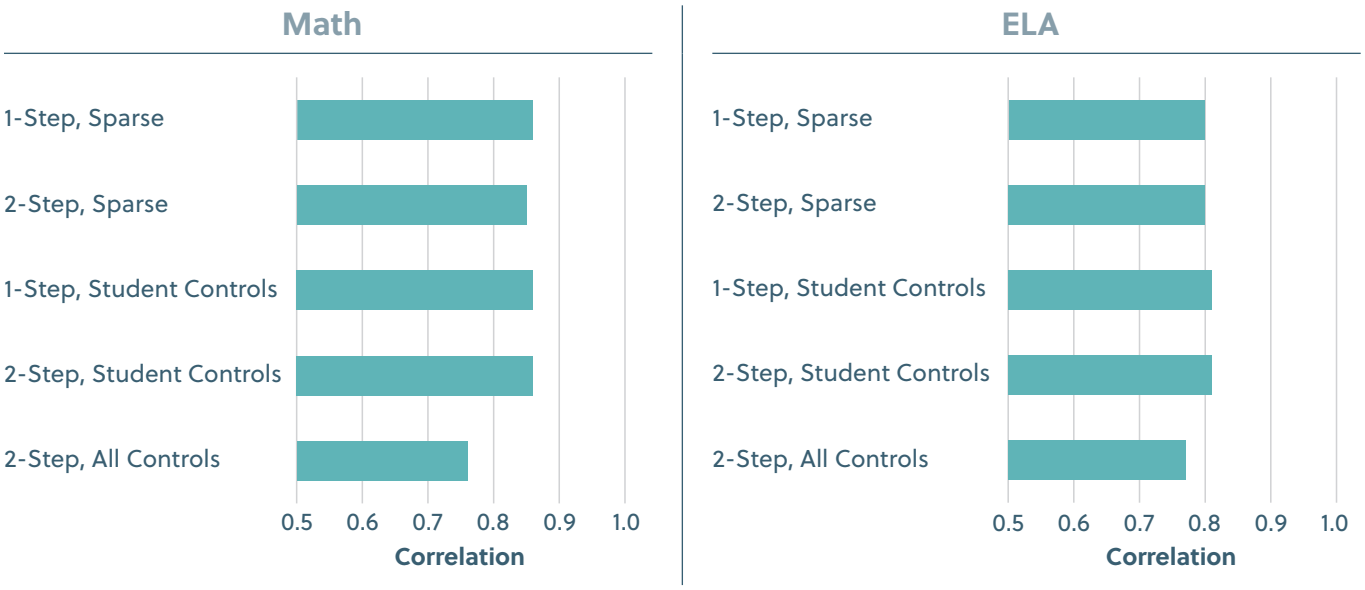
Two-year growth as a proxy for one-year growth

First, we consider the potential for using growth from 2019 to 2021 to proxy for growth from 2020 to 2021 (with the latter being unobserved). We simulate this situation with our data by correlating gap-year growth estimates with growth estimates from 2018 to 2019. We expect weaker correlations in this scenario because the growth metrics no longer span the same time periods. Put another way, additional differences in these estimates should emerge because one estimate captures growth over a two-year period and the other captures growth over a one-year period.

We find that gap-year growth for districts and schools is strongly correlated with one-year growth during the contemporary year, but the timespan inconsistency lowers the correlations relative to the comparisons in the main report. Figures A1 and A2 show correlations that correspond to the correlations shown in Figures 1 and 7 (for districts and schools, respectively) for this exercise. The correlations fall as expected—from about 0.90 into the range of 0.77 to 0.86 for districts and from about 0.87 into the range of 0.74 to 0.86 for schools. Note that the correlations fall the most for the estimates from the two-step, all-controls model. Although we do not elaborate in detail on this discrepancy, a likely explanation is that the less comprehensive models produce estimates that are biased due to uncontrolled district and school circumstances.[20] Any such bias would persist to some degree across years, which would inflate the correlations of these time-mismatched values relative to the two-step, all-controls model, where the bias will be smaller (or null).

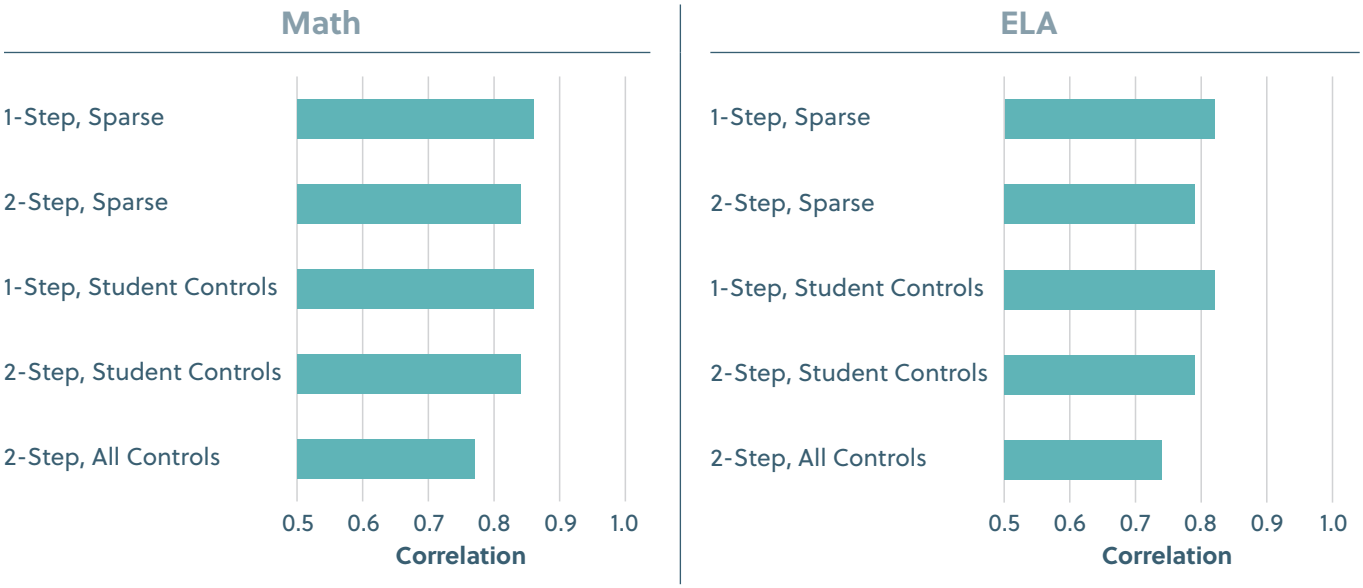
We provide these estimates as a point of information but do not explore this scenario in more detail because it does not have a strong policy rationale. To elaborate briefly, there are no testing data in 2020, which means that districts and schools were not rated on growth from 2019 to 2020. Rather than ignore that year altogether—which is implicit in this scenario—it seems appropriate to calculate growth over the full two-year timespan since the previous set of growth estimates were available. This, combined with the fact that we can most credibly estimate growth over the two-year span from 2019 to 2021, suggests that the time-consistent application of the gap-year model is most useful for policy.

Figure A1. Gap-year growth for districts is strongly correlated with one-year growth during the most recent year, but the timespan inconsistency lowers the correlations relative to the preceding comparisons.



Notes. The reported correlations are between growth estimated from spring 2017 to spring 2019 and growth estimated from spring 2018 to spring 2019. This informs a contemporary scenario where growth from spring 2019 to spring 2021 is used to proxy for growth from spring 2020 to spring 2021. For details on the models, see *Methodology*.

Figure A2. Gap-year growth for schools is strongly correlated with one-year growth during the most recent year, but the timespan inconsistency lowers the correlations relative to the preceding comparisons.

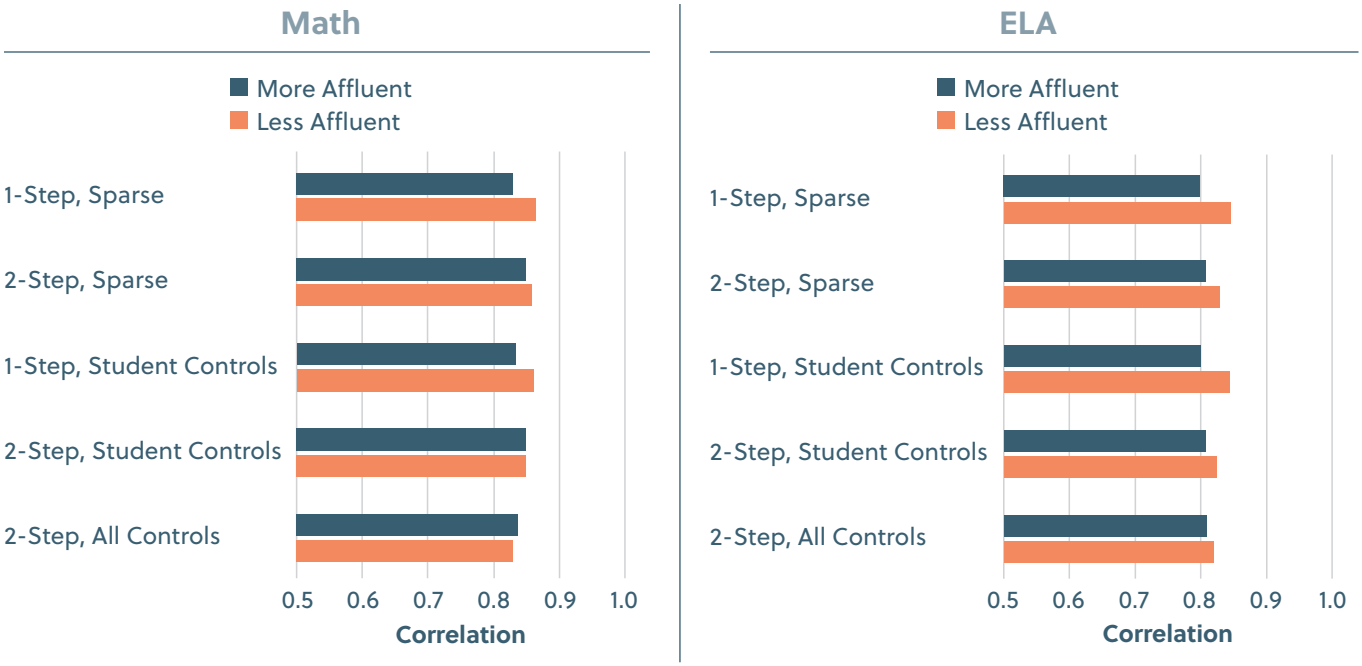


Notes. The reported correlations are between growth estimated from spring 2017 to spring 2019 and growth estimated from spring 2018 to spring 2019. This informs a contemporary scenario where growth from spring 2019 to spring 2021 is used to proxy for growth from spring 2020 to spring 2021. For details on the models, see *Methodology*.

Student subgroup analysis for schools

In the body of the report, we find that estimates of student subgroup growth within districts based on gap-year data for students of higher- and lower-socioeconomic status are highly correlated with growth estimates based on all of the data (Figure 2). Here, we replicate this analysis for schools, finding a similar result (Figure A3).

Figure A3. For students of varying socioeconomic status, the correlations of the school-level growth estimates with and without the gap-year data censoring remain high.



Notes. We refer to students coded as ineligible for free or reduced-price lunch as "more affluent" and those who are coded as eligible as "less affluent." For details on the models, see *Methodology*.

Endnotes

1 Guidance from the U.S. Department of Education indicates no plans to issue waivers from federal testing requirements for spring 2021. Andrew Ujifusa, "Betsy DeVos Tells States Not to Expect Waivers From Annual Tests," *Politics K–12* (blog), *Education Week*, September 3, 2020, http://blogs.edweek.org/edweek/campaign-k-12/2020/09/betsy-devos-annual-tests-not-expect-waivers.html.

2 Throughout our study, we assess the accuracy of gap-year growth estimates in terms of their ability to match growth that would be estimated if there was not a gap year. This approach relies on the assumption that under normal testing conditions, growth measures are useful indicators of district and school effectiveness. It is outside of the scope of our study to interrogate this assumption; however, there is a large research literature that provides evidence in its favor. This is consistent with the common use of growth modeling in education research and policy applications today.

3 Cory Koedel, Kata Mihaly, and Jonah E. Rockoff, "Value-added modeling: A review," *Economics of Education Review* 47 (August 2015): 180–195, doi:10.1016/j.econedurev.2015.01.006.

4 The precise specifications we use, along with technical details, are available in the following publication: Ishtiaque Fazlul, Cory Koedel, Eric Parsons, and Cheng Qian, "Estimating test-score growth with a gap year in the data" (unpublished manuscript, 2021).

5 There are two somewhat common growth-model approaches not directly covered by our analysis: student growth percentiles and EVAAS. With regard to the former, although we do not estimate student growth percentiles directly, a linear model using similar information produces similar results, which implies that our results using the sparse VAMs should be a reasonable approximation of results that would be obtained using student growth percentiles; see Mark Ehlert, Cory Koedel, Eric Parsons, and Michael Podgursky, "Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter?" *Educational Policy* 30, no. 3 (2016): 465–500, doi:10.1177/0895904814557593. We also do not evaluate EVAAS, which is a proprietary growth model owned by the SAS Institute. SAS's protectionary stance over EVAAS has been problematic in policy applications and was a sticking point in a recent lawsuit in the Houston Independent School District.

6 See Koedel, Mihaly, and Rockoff, "Value-added modeling." Noting this, our comparative results are substantively similar if we do not shrink the estimates.

7 See Ehlert, Koedel, Parsons, and Podgursky, "Selecting Growth Measures"; Dan Goldhaber, Joe Walch, and Brian Gabele, "Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments," *Statistics and Public Policy* 1, no. 1 (2014): 28–39, doi:10.1080/2330443X.2013.856169; Cassandra M. Guarino, Mark D. Reckase, and Jeffrey M. Wooldridge, "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy* 10, no. 1 (Winter 2015): 117–56, doi:10.1162/EDFP_a_00153; Cassandra M. Guarino, Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge, "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures," *Journal of Educational and Behavioral Statistics* 40, no. 2 (2015): 190–222, doi:10.3102/1076998615574771; Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger, *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment* (Seattle, WA: Bill and Melinda Gates Foundation, January 2013), https://files.eric.ed.gov/fulltext/ED540959.pdf; Koedel, Mihaly, and Rockoff, "Value-added modeling,"; and Eric Parsons, Cory Koedel, and Li Tan, "Accounting for Student Disadvantage in Value-Added Models," *Journal of Educational and Behavioral Statistics* 44, no. 2 (2018): 144–79, doi:10.3102/1076998618803889.

8 None of these models match the specification of the growth model in Missouri, although the two-step, all-controls model is the most similar to Missouri's model.

9 On the order of about 1 to 2 percent of schools and districts and about 0.1 percent of students are omitted (note that the omitted schools and districts enroll few students by definition), depending on the model and comparison.

10 Ben Backes, et al., "The Common Core Conundrum: To What Extent Should We Worry that Changes to Assessments and Standards Will Affect Test-Based Measures of Teacher Performance?" *Economics of Education Review* 62 (October 2017): 48–65, doi:10.1016/j.econedurev.2017.10.004.

11 For example, the predictive value of prior achievement as the testing regime changes is highly stable at the student level.

12 We elaborate on these factors in Fazlul, Koedel, Parsons, and Qian, "Estimating test-score growth."

13 For brevity, we only show results for the two-step, all-controls model. The results for the other models are very similar and their presentation is redundant (these results are available in Fazlul, Koedel, Parsons, and Qian, "Estimating test-score growth.").

14 Socioeconomic status is based on an identifier for whether a student is eligible for free or reduced-price lunch (and we note that these data are imperfect in Missouri because they are censored due to community eligibility). We refer to students coded as ineligible as "more affluent" and those who are coded as eligible as "less affluent."

15 The variable that drives most of the explanatory power in each model is the average same-subject test score from the base year—i.e., the 2017 score. Districts with higher lagged test scores experience larger ranking declines, all else equal, in the first four models when we shift to the gap-year model; in the two-step, all-controls model, they experience very modest ranking increases. We have attempted to understand theoretically why this is happening in the models, but we cannot generate a theoretically grounded *ex ante* prediction of this result. We are left to conclude that it is an *ex post* realization of results based on differences in how the models predict current achievement using lagged versus twice-lagged achievement that may not replicate in other circumstances. The influence is subdued in the two-step, all-controls model because of its rich controls for lagged achievement at the school and district levels, highlighting an advantage of that approach.

16 In an undated report, The SAS Institute gives analogous correlations for growth estimates for districts (and schools and teachers) from its proprietary TVAAS Multivariate Response Model (MRM). SAS reports much higher correlations—0.99 for districts and schools—using a similar research design that compares data conditions that differ by whether there is a gap year in testing. The SAS Institute report does not address the issue of cohort overlap, nor does it directly investigate student mobility. We are surprised, and frankly skeptical, that TVAAS growth estimates are so impervious to the missing year of data but are unable to explore the discrepancy in results given the propriety nature of their model.

17 We use the 2018 test data to assign students to a district in 2018. With a true gap year, test data would be unavailable to assign the 2018 districts, but this could be achieved using enrollment records instead.

18 The fraction of test-taking students who are structural movers is much higher than the fraction of all students because structural moves are uncommon in non-test-taking grades (i.e., grades K through 2 and grades 9 through 12).

19 If we do not observe full, self-contained cohorts within schools, the only way to estimate test-score growth for each school would be to rely on complex and assumptive statistical tools to separate out the contributions of different schools to each student's growth during the extended period. We do not view this as a feasible policy alternative.

20 Parsons, Koedel, and Tan, "Accounting for Student Disadvantage."