Teacher Effectiveness in Remote Instruction

M. Cade Lawson
Tim R. Sass

September 2024

WORKING PAPER No. 308-0924

CALDER
National Center for Analysis of Longitudinal Data in Education Research

AIR®

Teacher Effectiveness in Remote Instruction

M. Cade Lawson
Georgia State University

Tim R. Sass
Georgia State University

Contents

Contents i
Acknowledgments ii
Abstract iii
1. Introduction 1
2. Data 2
3. Methods 3
4. Results 6
5. Conclusion 16
References 18
Tables and Figures 19
Appendix 30

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A230400 to Georgia State University, and by the Center for Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations (see https://caldercenter.org/about-calder).

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

Teacher Effectiveness in Remote Instruction
M. Cade Lawson & Tim R. Sass
CALDER Working Paper No. 308-0924
September 2024

Abstract

The effect of remote learning on student performance has been a frequent topic of research and discussion in the aftermath of the COVID-19 pandemic, yet little is known about the impact of remote instruction on the performance of teachers. This study documents how relative effectiveness of teachers changed when moving from in-person to remote instruction and analyzes the characteristics of teachers associated with greater relative effectiveness during remote instruction. Using matched student/teacher-level data from three large metro-Atlanta school districts, we estimate teacher value-added models to measure the association between teacher characteristics and a teacher's relative contribution to test score growth before and during the period of virtual instruction in the 2020-21 school year. We find evidence of increased variation in overall teacher effectiveness during remote instruction. Results are driven by veteran teachers, who appear relatively more effective in virtual instruction than their less-experienced peers, and by the very best in-person teachers, some of which experience large declines in relative effectiveness when shifting to remote instruction.

1. Introduction

Despite the poor performance of students learning virtually during the COVID-19 pandemic, it appears that remote instruction will become a permanent part of the K-12 education landscape in the U.S. So-called “virtual learning days” have increasingly been used to substitute for in-person instruction during inclement weather, while greater familiarity with online delivery platforms has made remote instruction a more plausible solution for addressing shortages in “hard-to-staff” subject areas and for offering more advanced courses.

Though there is a large literature on teacher effectiveness during traditional in-person instruction, little quantitative evidence has emerged to characterize teacher effectiveness during remote instruction. In this study, we use longitudinal student-level achievement data from three large school districts in the metro-Atlanta area to investigate the characteristics of effective remote teachers. We find that variation in teacher effectiveness increased during remote instruction, with results driven by large losses in relative effectiveness among some of the best in-person teachers and modest gains for veteran teachers relative to their less-experienced colleagues.

The results of this work can be used to inform state and district policy on issues such as selection of teachers to provide virtual instruction, instructional delivery methods, and in-service training on technology use. It also carries important equity implications, considering evidence on the role remote instruction played in widening existing achievement gaps during the pandemic (Goldhaber et al., 2023).

1.1 Prior Research on Effective Teaching and Remote Instruction

To date, studies attempting to characterize effective remote instruction have relied on survey evidence collected from online courses at the post-secondary level with a particular focus on pedagogical practices or self-reported factors. For example, some of the earliest evidence suggests a strong correlation between certain teacher practices (such as use of examples in class and adapting to student needs) and perceived teacher quality (Young, 2006), and between teaching performance and personality traits (Holmes et al., 2015). Similarly, qualitative evidence from interviews of academics in Australia and New Zealand reports a set of pedagogical "attributes" most associated with effective online instruction (Rose, 2018). In contrast with these methodologies, we provide evidence on teacher effectiveness in remote instruction using longitudinal administrative data from students at the K-12 level and focus on teacher characteristics rather than pedagogical choices in characterizing effective instruction. This allows us to observe the extent to which remote instruction may interact with established relationships from the in-person teacher effectiveness literature, some of which have been shown to differ during the period of virtual instruction (North Carolina Dept. of Public Instruction, 2023).

2. Data

This study uses data from three large metro-Atlanta school districts. It includes student-level test scores on nationally normed formative assessments, including fall and winter test scores in two districts and fall, winter, and spring test scores in a third district, in each of math and reading. The test scores are normalized by grade and year based on pre-pandemic national means and standard deviations. In addition, the data contain student demographic variables such as free/reduced-price lunch status, disability status, race/ethnicity, and gender, as well as teacher characteristics such as gender, race/ethnicity, and years of experience. We use data on periods of virtual versus face-to-face instruction in each district during the pandemic to construct indicators for semesters when teachers were instructing remotely. The panel structure of the data allows us to follow both teachers and students over time during the period of observation. To further describe the extent of our data, we provide the aggregate counts of student and teacher data used in our analysis by district and school year in Table 1.

Districts 1 and 2 operated with universal remote instruction during the fall semester of school year (SY) 2020-21. Students in District 2 returned to in-person learning in the middle of the Spring 2021 semester (SY 2020-21). Students in District 1 switched back to in-person later, with elementary students returning to in-person learning during the Spring 2021 semester and middle and high-school students returning to in-person instruction at the start of SY 2021-22. District 3 had a multi-phase return to in-person learning during Fall 2020, beginning with universal remote delivery for all students and gradually ramping up in-person learning to one day per week, then two days per week, and eventually five days per week in-person. We use daily student-level attendance-by-learning-mode data to calculate the share of teaching done remotely by each teacher in District 3 during this hybrid semester. We focus our analysis on teacher effectiveness during two key semesters: Fall 2019, the last full semester before remote instruction began, and Fall 2020, the semester during which districts operated under fully-remote (Districts 1 and 2) or partially-remote (District 3) delivery.

3. Methods

We quantify teacher effectiveness by estimating a series of value-added models, a class of regression model which assesses a teacher's ability to generate test score growth among students. Our modelling approach allows us to compare teachers within-district, subject, and semester, giving each teacher a value-added estimate for each subject they teach for each semester they are observed in our data. We estimate teacher value added over the fall semester by predicting a student's winter (mid-year) test score as a function of their fall (beginning-of-year) test score plus other student demographic and class controls. By controlling for appropriate factors, the difference between the student's predicted and actual score can be attributed to the teacher. Such teacher value-added estimates have been shown to be unbiased measures of teacher quality when evaluated with external data (Chetty, Friedman & Rockoff, 2014).

3.1 Empirical Model

We estimate value-added using the following functional form, where our specific measure of value-added is given by the teacher fixed effect μ_s:

$$Y_{(w)ijs} = \beta_0 + Y_{(f)ijs}\beta_1 + X_{is}\beta_3 + Z_{ijs}\beta_4 + \mu_s + \epsilon_{ijs} \quad (1)$$

Equation 1 models the winter (end of semester) test score, $Y_{(w)ijst}$, for student i in subject $j \in$ {reading, math} with teacher s as a function of the prior test score, and student, teacher, and class characteristics. $Y_{(f)ijs}$ is a vector that includes the student's prior (Fall) test score and its square, to capture possible nonlinearities in the relationship between current and past performance. X_{is} the vector of student-level information, including race/ethnicity, disability status, free/reduced-price meal status, gender, English language learner status, and the number of instructional days between exams. Z_{ijs} represents class size, and μ_s is our final parameter of interest: the value-added for teacher s, measured in terms of test score standard deviations centered around the average teacher in the relevant subject, district, and semester.[1]

3.2 Addressing Noise in Value-Added Estimates

Value-added, especially when estimated with a relatively low number of student test scores for a given teacher, can be a noisy measure of teacher effectiveness. We take several steps to mitigate this issue. First, we estimate our models using only test scores corresponding to students who received instruction from a single teacher in a given subject and who were in classes with just one instructor of record. This ensures we properly attribute test score growth to each teacher, as it is difficult to disentangle proportions of score growth attributable to multiple

[1] Our model excludes any observable teacher characteristics. Consequently, the value-added estimates represent relative performance across all teachers, which may embody differences due to teacher experience or other characteristics that are associated with teacher performance.

teachers for a single student in the same semester. We restrict the sample to students who began the course in August or September to ensure that sufficient time was spent being taught by the teacher. In addition, we only consider test scores coming from classes with at least 10 total scores in a given semester. Classes with fewer students could reflect data reporting errors or systematic non-testing of students along some unobservable dimension. Given the substantial imprecision in value-added estimates for teachers with fewer than 10 students (McCaffrey et al., 2009), we do not calculate a value-added estimate for any teacher who has fewer than 10 usable test scores within a semester and subject area.

We use empirical Bayes shrinkage to further reduce variability in our estimates of teacher effectiveness. This approach addresses underlying noise by shrinking value-added estimates for teachers towards the sample mean (which is zero, in the case of the value-added model). A teacher with very few underlying test scores is moved more towards the mean than a teacher with many observed scores under the assumption that true teacher fixed effects are distributed normally with mean zero. In the absence of any teacher-specific evidence on performance, the best prediction of teacher effectiveness is therefore zero. The more confidence one has in a teacher's impact on student test scores, the greater the weight placed on the estimate of their individual performance and the less "shrinkage" is done. To deal with issues posed by the arbitrary omission of a reference teacher in usual fixed effects modeling, we estimate Equation 1 using the *felsdvregdm* routine in STATA. This routine is designed to impose a sum-to-zero constraint on fixed effects within a reference group, allowing us to appropriately apply EB shrinkage post-estimation (Mihaly, McCaffrey, & Sass 2010).

3.3 Interpreting Value-Added During Remote Instruction

In each of our models, we center our value-added estimates on the average K-8 teacher within a given semester, district, and subject. As such, our estimates can be interpreted as the additional test score growth associated with a particular teacher relative to the average K-8 teacher in her district who is teaching the same subject in the same semester.

For Districts 1 and 2, we compare teacher value added in the fall semester of SY 2019-20 (when there was universal in-person instruction), with teacher value added in the fall semester of SY 2020-21 (when all students learned remotely). Of course, the two periods differ not only in the instructional mode employed, but also in the existence of the COVID-19 pandemic. To the extent that the effect of the pandemic on student learning (independent of instructional mode) is captured by baseline test scores and observable student characteristics, this does not pose a problem for the analysis. Further, given that value added is a relative measure of teacher performance, any impacts of the pandemic that affect all teachers equally (conditional on student characteristics and prior scores), will drop out when centering on the average teacher. Likewise, if the pandemic affected teachers randomly, the effect of the pandemic would be captured by the error term in Equation (1).

When assessing the relationship between teacher characteristics (e.g. experience) and value added in remote instruction, the effects of the pandemic and instructional mode would only be conflated if the pandemic systematically affected teacher performance in a non-uniform way that is correlated with teacher experience.

4. Results

We begin with an analysis of Districts 1 and 2, which both had fully remote instruction throughout the Fall 2020 semester. To discern whether there are systematic differences in the determinants of teacher effectiveness between in-person and remote instruction, we first calculate teacher value-added for teachers in Districts 1 and 2 by subject and year and correlate these annual measures with one-another. If teacher value-added from the last semester of in-person instruction (Fall of SY 2019-20) and the first semester of remote instruction (Fall of SY 2020-21) are more weakly correlated than are estimates from any two preceding contiguous face-to-face instruction semesters, there are likely underlying differences in the characteristics of (relatively) effective teachers during remote instruction.[2] We present these results in Table 2.

The diagonal entries in Table 2 show the year-to-year pairwise correlations in a teacher's value-added score. During in-person instruction the year-to-year correlation in value added is fairly consistent, ranging from 0.28 to 0.35. These correlation coefficients are in line with established norms on intertemporal variability from the value-added literature (McCaffrey et al., 2009). However, as shown in the lower-right-hand corner of Table 2, the correlation between teacher value-added during remote instruction (Fall 2020) and value-added in the previous year (Fall 2019) is much weaker; the pairwise correlation is only 0.11. In other words, there is more shifting in relative teacher performance when moving to remote instruction than one would expect from normal year-to-year changes in relative performance of teachers, holding instructional mode constant. To allay concerns that these results are driven by compositional change (such as teacher turnover), we present correlations for the sub-sample of teachers who are observed in all five semesters in Appendix Table A1. The correlations from this consistent set of teachers reveals similar disparities between relative performance during remote instruction relative to in-person teaching.

[2] It is important to keep in mind that our value-added measure is a gauge of relative performance. It is likely that, on average, teachers were less effective in promoting student learning during the pandemic for a variety of reasons. We are arguing here that the reduction in year-to-year correlation of value-added is signaling a re-arranging of teacher relative performance during remote instruction.

4.1 The Teacher Effectiveness Distribution by Instructional Mode

The finding that the relative performance of teachers changes with the instructional mode begs the question of which teachers are relatively better at remote instruction. To begin addressing that question, we first compare the overall distributions of teacher value-added before and during remote instruction among teachers observed during both periods. Because we estimate value-added separately by district, subject, and semester, an individual value-added estimate reflects that teacher's ability to generate test score growth relative to the average teacher in that same semester (who has a value-added estimate of zero by construction). As such, we can compare the variances of the value-added distributions from Fall 2019 (in-person) and Fall 2020 (remote) to understand whether teacher performance had a larger "spread" during remote instruction. The results of this exercise are shown in Figure 1.

Figure 1 indicates that there was increased variation in teacher value-added during remote instruction, especially at the left tail: while teaching remotely, a number of teachers had value-added estimates lower than the minimum of all teacher value-added estimates from in-person instruction. We present evidence that the difference in value-added variance between in-person and remote instruction is statistically different from zero using Levene's test of equality of variances in Table A2. These differences persist when breaking down the distributions by subject (math vs. reading) and by grade level (elementary vs. middle). We present these results in Appendix Figures A1 and A2, respectively, with corresponding statistical tests of equality of variance in Table A3.

4.2 Descriptive Analysis of the Teacher Effectiveness Distribution by Instructional Mode and Teacher Characteristics

Teacher Experience

One of the few observable characteristics of teachers that is consistently associated with effectiveness during in-person instruction is teacher experience. While teacher effectiveness generally increases throughout a teacher's career, the greatest increase in teacher value-added from experience occurs during the first few years on the job (Harris & Sass, 2011; Kini & Podolsky, 2016). If the skills or practices that improve with experience, such as classroom management or well-developed lesson plans, are relatively more important in remote instruction, we could see a positive relationship between teacher experience and effectiveness in remote instruction. Alternatively, it could be that the requisite skills to excel in remote instruction are very different than those required for in-person instruction. For example, the ability to utilize digital instructional materials may be particularly important in remote instruction. If younger (and less experienced) teachers are more facile with the digital tools used for virtual instruction, we could observe an inverse relationship between teacher experience and relative performance in remote instruction. To examine the relationship between experience and differences in teacher effectiveness across instructional modes, we compare changes in the distribution of value-added for three sub-groups of teachers: early, mid, and late-career teachers with 0-5, 6-15, and greater than 15 years of experience, respectively. We report these results in Figure 2.

The changes in the teacher effectiveness across learning modes are strikingly similar across the three teacher experience categories. For each group, the move to remote instruction is associated with an increase in the variance of teacher effectiveness, with increased density of teachers in the left tail of the distribution (i.e. a greater proportion of teachers with very low value-added relative to the average). There are some subtle differences across the ability groupings. It appears that a marginally greater proportion of early-career teachers demonstrated extremely low value-added in remote instruction than did mid-career and late-career teachers. A slightly larger proportion of late-career teachers had relatively higher value-added in remote instruction than did early-career or mid-career teachers. However, these visual comparisons do not clearly reveal a strong relationship between experience and effectiveness in remote instruction aside from modest differences in average effectiveness between more and less-experienced teachers. Table 3 presents a statistical comparison of variances for the distributions in Figure 2.

Levene's test for equality of variances finds significant differences in variance within each teacher experience category between in-person and remote teaching, but no joint differences across experience categories within-semester regardless of learning modality.

Initial Effectiveness in Remote Instruction

Given there are many unobserved teacher characteristics that drive teacher effectiveness, we consider the relationship between teacher effectiveness in remote instruction (which is determined both by observed characteristics we can observe, like experience, and unobserved factors, like personality characteristics) and the change in effectiveness when moving to remote instruction. We examine changes in effectiveness level by placing teachers into five groups based on the quintile of their in-person value-added estimate during Fall 2019. We then estimate each teacher's value-added quintile during remote instruction in Fall 2020 and tabulate the number of teachers who move between each pair of quintiles: from 1st quintile to 1st quintile, 1st to 2nd, etc. between successive school years. This approach allows us to check for abnormally large or small outflows of teachers from one quintile to another, which could provide suggestive evidence on whether effective in-person teachers are still effective as remote instructors.

Even between any two in-person semesters, value-added measures are expected to fluctuate. For example, some year-to-year changes could reflect teachers getting an unusually strong (or unusually weak) set of students one year and a more-typical set of students the next, causing the teacher's measured value-added to fall (or rise) from one year to the next. To mitigate the possibility of mistaking these normal year-over-year fluctuations in effectiveness for differences during remote instruction, we include a baseline measure calculated by comparing a teacher's effectiveness quintile during Fall 2018 (an in-person semester) with her effectiveness quintile during Fall 2019 (another in-person semester). We display the results of this exercise in Table 4.

Table 4 tracks the rate of quintile "agreement" or "stability" for teachers across semesters. Panel (A) tracks quintile agreement between Fall 2019 and Fall 2020, the transition from in-person to remote instruction, and Panel (B) provides baseline agreement estimates from the transition from Fall 2018 to Fall 2019 (two in-person semesters). The lower agreement values along the diagonal in Panel (A) compared to those in the diagonal in Panel (B) emphasize the previously discussed increased variability of value-added estimates during remote instruction: a lower percentage of teachers stay in the same effectiveness quintile during remote instruction than between the baseline in-person years. Though few clear trends emerge to indicate where teachers who now change quintiles are more likely to end up, Panel (A) provides some suggestive evidence that the most effective teachers (in the 4^{th} and 5^{th} quintiles of initial effectiveness) were more likely to fall to very low quintiles during remote instruction: nearly 37% of teachers who were in the 5^{th} quintile of effectiveness in Fall 2019 fell to the 1^{st} or 2^{nd} quintile during remote instruction, compared to just 23% during the baseline period.

We provide an alternative view of these quintile changes in Figure 3, which tracks the starting and ending quintiles for all teachers over the two pairs of years. Each panel tracks the percent of all teachers who transition from the initial effectiveness quintile indicated in the title to the final quintile indicated on the horizontal axis. The red bars indicate the percent of all teachers making such a quintile change between in-person and remote instruction (Fall 2019 to Fall 2020), while the blue bars indicate the percent of teachers making the same type of quintile change between two in-person semesters (Fall 2018 to Fall 2019). Thus, for example, in the first panel the height of the second blue bar indicates that roughly four percent of teachers were in the bottom quintile of teachers in SY 2018-19 and in the second quintile of the teacher value-added distribution in SY 2019-2020. The height of the second red bar in that same panel shows that approximately 4.5 percent of all teachers were in the first quintile of the value-added distribution in SY 2019-20 and moved to the second quintile in the value-added distribution for SY 2020-21.

Examining just teachers who were in the fifth quintile in SY 2020, the bottom-right panel, we see that they were substantially more likely to fall from the 5th to the 1st quintile during the in-person-to-remote period than the in-person-to-in-person control period. The same pattern holds for teachers initially in the 4th quintile, who were more likely to fall from the 4th to the 1st quintile during remote instruction than in a typical pair of in-person instruction years. Beyond these differences, a large proportion of teachers remain in the same quintile of the teacher effectiveness distribution after transitioning to remote instruction. Summing the heights of the five red bars designating no change in quintile rank between SY 2019-20 and SY 2020-21, we find that about one-fourth of teachers ranked in the same quintile in both years.

While Table 3 and Figure 3 provide suggestive evidence that effective in-person teachers generally remained relatively effective during remote instruction, we emphasize that value-added estimates are noisy measures of teacher performance due to measurement error in student test scores. Thus, changes in value added can be due in part to random fluctuations in test scores and one should interpret the quintile changes in value-added with caution.

4.3 Teacher Effectiveness Under Partially Remote Delivery

While Districts 1 and 2 both operated with fully remote instruction during Fall 2020, District 3 had a phased return to in-person instruction over the course of the Fall 2020 semester that saw students attend school virtually at first and then for one, two, and eventually five days per week in-person based on local public health trends. To quantify the amount of remote instruction done by each teacher in District 3 during Fall 2020, we use student-level attendance data to calculate the share of days a student attended school virtually and average this value for each teacher. The average teacher in District 3 had a remote teaching share of 0.633, implying that students attended the teacher's class remotely 63% of the time on average during Fall 2020. We present the distribution of these teaching shares in remote instruction in Appendix Figure A3.

We conduct similar analyses, as were performed for Districts 1 and 2, to understand the extent to which partially in-person delivery may have disrupted trends in teacher effectiveness. As was the case with fully remote instructional delivery in Districts 1 and 2, teachers in District 3 display higher variance in effectiveness during the partially remote period. Figure 4 overlays the teacher effectiveness distributions in District 3 during in-person and partially remote instruction.

The relatively low correlation between years of experience and effectiveness as an online instructor is more pronounced in the partially remote setting of District 3. Whereas in Districts 1 and 2 there was some evidence that veteran teachers outperformed their peers as remote instructors, evidence from District 3 implies no substantial heterogeneity in teacher performance during remote instruction from differences in teacher experience. If anything, it appears that existing discrepancies in effectiveness along this dimension were reduced during remote instruction. Figure 5 displays these results. The leftward shift of the distribution for veteran teachers after the onset of virtual instruction brings them roughly in line with less experienced peers, despite an advantage during in-person instruction. These results are demonstrated empirically with a statistical comparison of variances in Table A4.

We also reproduce our results on differences by initial effectiveness quintile in the context of District 3. These results, reported in Table 4, are quite similar to analogous results from Districts 1 and 2 presented in Table 3. Teachers in the highest initial performance quintile were once again more likely to fall to the 1st and 2nd quintiles during remote instruction as compared to the baseline, though the relationship is more muted in District 3. The most notable difference is that rates of remaining in the same quintile year-over-year are not systematically lower during remote instruction in District 3 as they are Districts 1 and 2. This result may be driven by the relatively smaller sample of teachers observed teaching during both Fall 2018 and Fall 2019 (the two baseline semesters) in District 3 as a result of its formative assessment rollout. We present analogous results to Figure 3 for the setting of District 3 in Appendix Figure A4, which again reveals similar trends in teacher quintile changes.

4.4 Disentangling Remote Learning and Pandemic Effects

While our strategy for estimating teacher value-added allows for a common shock from the COVID-19 pandemic to affect all teachers equally without being reflected in our results, it remains a possibility that some patterns we observe in our analysis are associated with pandemic effects rather than the transition to virtual instruction. We address that possibility in this section by studying two specific time periods in our data – one in District 2 and one in District 3 – which may allow us to disentangle pandemic from remote learning effects.

We begin by contrasting the distributions of teacher effectiveness during Fall 2020 and Spring 2021 in District 2. Because this district administers formative assessments in the fall, winter, and spring, we are able to estimate value-added using both fall-to-winter and winter-to-spring test score growth. This allows us to make comparisons within-academic year by constructing separate Fall and Spring value-added measures. The Fall 2020 semester was fully remote in District 2, while the Spring 2021 semester featured a mix of in-person and remote instruction. This setting is unique because the pandemic was ongoing during both semesters of SY 2021, while universal remote instruction was only in place during Fall 2020. As such, we would expect to see no substantial change in the teacher effectiveness distribution during Spring 2021 in the event that the increased variation in teacher effectiveness from Fall 2020 was driven by the pandemic factors rather than those related to remote teaching. We present the results of this exercise in Figure 6.

Figure 6 reveals a tightening of the teacher effectiveness distribution during Spring 2021, a reversion of the increased variance observed between Fall 2019 and Fall 2020 in Figure 1. Because the pandemic was ongoing during both semesters of SY 2021, this provides suggestive evidence that the widening of the teacher effectiveness distribution was driven by the transition to online learning rather than pandemic-related factors. We provide empirical support for the tightening of the distribution using an equality-of-variances test in Table A5.

In addition, we study the partially-remote teaching environment during Fall 2020 in District 3 to provide additional evidence on the extent to which our results are driven by online instruction vs. pandemic effects. We divide the sample of teachers for whom we estimate value-added into subsamples with above-median and below-median shares of time spent teaching remotely during Fall 2020 and then examine how the distributions of the two groups change compared to their pre-pandemic baselines from Fall 2019. If remote teaching itself is driving the increased spread of teacher value-added during remote instruction, one would expect teachers with higher shares of remote instruction during Fall 2020 to have higher variance in value-added compared to their Fall 2019 baseline than those with low shares of remote instruction. This inclusion of the Fall 2019 baseline for each group also partially controls for non-random sorting of teachers into high or low shares of remote instruction if, for example, administrators systematically assigned their best teachers to do more remote teaching. We display the results of this exercise in Figure 7.

Figure 7 does not reveal a substantial difference in the distribution of value-added for teachers with above-median remote teaching shares (solid blue line) and those with below-median remote teaching shares (dashed blue line) during Fall 2020, a result confirmed empirically using Levene's comparison of variances test in Table A6. While this analysis fails to provide strong evidence of a remote teaching effect, this null effect could be attributable to a lack of variation in exposure to remote instruction. The majority of District 3 teachers had remote teaching shares of 5-80 percent in Fall 2020 and no teachers taught only in-person (see Figure A3). Had there been greater variation in remote-instruction shares, it is possible that there would be a greater disparity between the value-added distributions of above-median and below-median remote-instruction teachers.

5. Conclusion

Teachers are the most important school-based input to a student's education, but previous research has not quantified how the relative effectiveness of teachers varies by instructional mode. We leverage the unplanned universal shift to remote instruction brought on by the COVID-19 pandemic to analyze the factors associated with differences in relative performance across instructional modes. To do this, we estimate a series of value-added models for three large school districts in metro-Atlanta and examine the ways teacher value-added changed during the period of universal remote instruction in Fall 2020. In aggregate, we find higher variance in value-added during remote instruction; many teachers were either much more or much less effective at generating test score growth compared to in-person instruction. We also show that, conditional on initial effectiveness, more experienced teachers had larger increases in value-added during remote instruction. This suggests that veteran teachers may have been able to "weather the storm" of online teaching better than their less experienced peers, perhaps due to unobserved differences such as better-developed lesson plans or classroom management skills.

While it may be a natural inclination to assign the "best" teachers to specialized or ad-hoc tasks like teaching a class remotely, our initial evidence suggests that the best in-person teachers saw large declines in effectiveness at a higher rate during remote instruction as compared to an in-person baseline. As such, it might make sense to prioritize other characteristics besides in-person effectiveness when staffing remote classes -- such as experience, which we find to be associated with marginally higher relative performance in remote teaching. More precise policy guidance, such as what training teachers need to excel in a virtual learning environment, would require a better understanding of the specific skills or traits that are associated with relative performance during remote instruction.

References

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. American Economic Review, 104(9), 2593-2632.

Holmes, C. P., Kirwan, J. R., Bova, M., & Belcher, T. (2015). An investigation of personality traits in relation to job performance of online instructors. Online Journal of Distance Learning Administration, 18(1), 2015.

Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., & Staiger, D. O. (2023). The educational consequences of remote and hybrid instruction during the pandemic. American Economic Review: Insights, 5(3), 377-392.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. Journal of Public Economics, 95(7-8), 798-812.

Kini, T., & Podolsky, A. (2016). Does Teaching Experience Increase Teacher Effectiveness? A Review of the Research. Learning Policy Institute.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education finance and Policy*, *4*(4), 572-606.

North Carolina Department of Public Instruction, Office of Learning Recovery and Acceleration. (2023). Examining Teacher and Principal Effectiveness and Longevity on Mitigating the Negative Impacts of the Pandemic. https://www.dpi.nc.gov/blog/2023/01/30/examining-teacher-and-principal-effectiveness-and-longevity-mitigating-negative-impacts-pandemic

Mihaly, K., McCaffrey, D. F., Lockwood, J. R., & Sass, T. R. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *The Stata Journal*, *10*(1), 82-103.

Rose, M. (2018). What are some key attributes of effective online teachers?. *Journal of Open, Flexible and Distance Learning*, *22*(2), 32-48.

Stuit, D., Austin, M. J., Berends, M., & Gerdeman, R. D. (2014). Comparing estimates of teacher value-added based on criterion-and norm-referenced tests. *US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest.*

Young, S. (2006). Student views of effective online teaching in higher education. *The American Journal of Distance Education*, *20*(2), 65-77.

Tables and Figures

Table 1. Counts of teachers and students by school year, district, and subject

			2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
District 1	Math	Teachers	--	--	790	184	303	580
		Students	--	--	23,619	5,467	9,217	15,078
	Reading	Teachers	--	--	568	171	330	645
		Students	--	--	19,889	5,221	10,218	17,231
District 2	Math	Teachers	1,426	1,271	1,025	1,174	605	1,216
		Students	36,948	35,801	28,325	32,268	10,048	30,633
	Reading	Teachers	1,408	1,342	1,105	1,255	662	1,241
		Students	34,902	36,090	27,782	32,880	11,466	29,436
District 3	Math	Teachers	--	47	1,023	1,058	942	1,200
		Students	--	945	26,456	21,877	20,075	27,783
	Reading	Teachers	--	490	544	988	938	1,206
		Students	--	13,069	15,336	19,128	17,861	25,087

Note: Counts describe the total number of teacher-subject and student-subject observations used for estimating fall-to-winter value-added models during each school year, in each district. The drop in student test scores for District 1 in SY 2019-20 reflects a transition between formative assessment vendors, which resulted in fewer students having both a fall and winter test score to be used in estimating value-added models, and less frequent reporting of class start dates for students. The low number of student test scores in District 3 in math during SY 2017-18 reflects the initial rollout of formative testing in the district, which began with reading.

Table 2. Annual pairwise teacher value-added correlations, Fall 2016 – Fall 2020

		Fall 2016	Fall 2017	Fall 2018	Fall 2019
In-Person Instruction	Fall 2016	1.000	--	--	--
	Fall 2017	0.346	1.000	--	--
	Fall 2018	0.268	0.287	1.000	--
	Fall 2019	0.360	0.352	0.276	1.000
Remote Instruction	Fall 2020	0.210	0.249	0.176	0.114

Note: Pairwise teacher value-added correlations above correspond to teachers in Districts 1 and 2, both of which had fully-remote delivery during Fall 2020.

Table 3. Comparison of variances by teacher experience and teaching mode, Districts 1 & 2

	Early-Career	Mid-Career	Late-Career	P-Value
Fall 2019 (In-Person)	0.022	0.026	0.023	0.754
Fall 2020 (Remote)	0.062	0.049	0.048	0.071
P-Value	0.000	0.000	0.000	

The table above displays the variance of value-added for teachers of each subject (indicated by each column header) during the time period indicated in each row. P-values represent Levene's test statistic for equality of variances, which is robust to non-normality. Tests of equality of variance reported at the end of each row compare distributions by teacher category within-years, while those at the end of each column compare distributions of the same teacher type across years.

Table 4. Teacher effectiveness quintile agreement, comparing successive Fall semesters, Districts 1 and 2

Fall 2019 (In-Person)	Fall 2020 (Remote): 1st	2nd	3rd	4th	5th
1st	**27.3%**	22.6%	20.7%	17.9%	12.3%
2nd	17.7%	**20.5%**	24.2%	23.6%	13.3%
3rd	17.7%	19.5%	**22.2%**	21.0%	20.5%
4th	21.2%	16.9%	17.7%	**18.5%**	25.1%
5th	16.2%	20.5%	15.2%	19.0%	**28.7%**

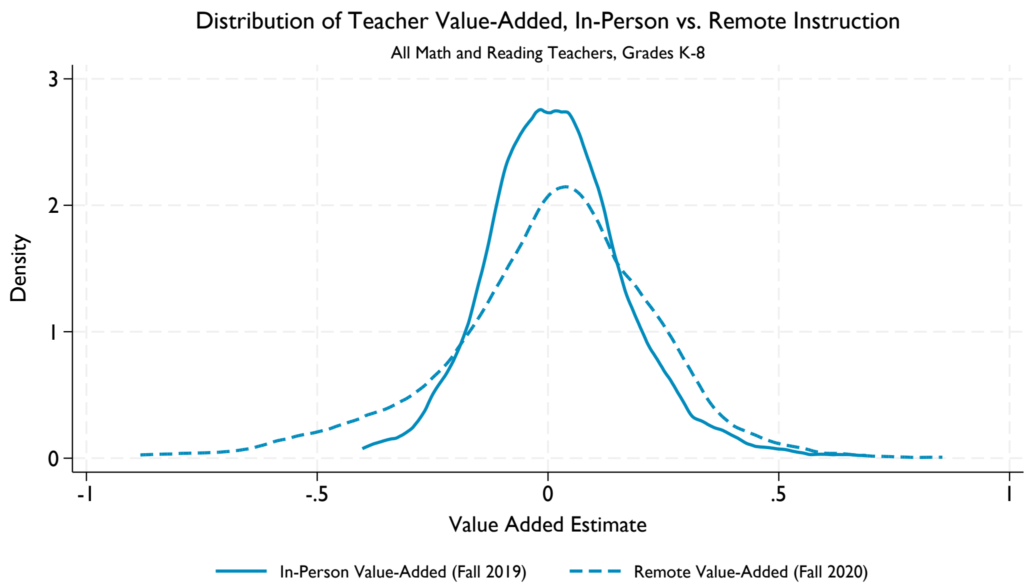
(A)

Fall 2018 (In-Person)	Fall 2019 (In-Person): 1st	2nd	3rd	4th	5th
1st	**33.0%**	25.0%	15.4%	12.3%	14.8%
2nd	23.3%	**24.0%**	22.0%	16.6%	14.4%
3rd	21.0%	17.5%	**23.0%**	23.7%	14.1%
4th	11.7%	21.4%	23.0%	**22.4%**	22.0%
5th	11.0%	12.0%	16.7%	25.0%	**34.8%**

(B)

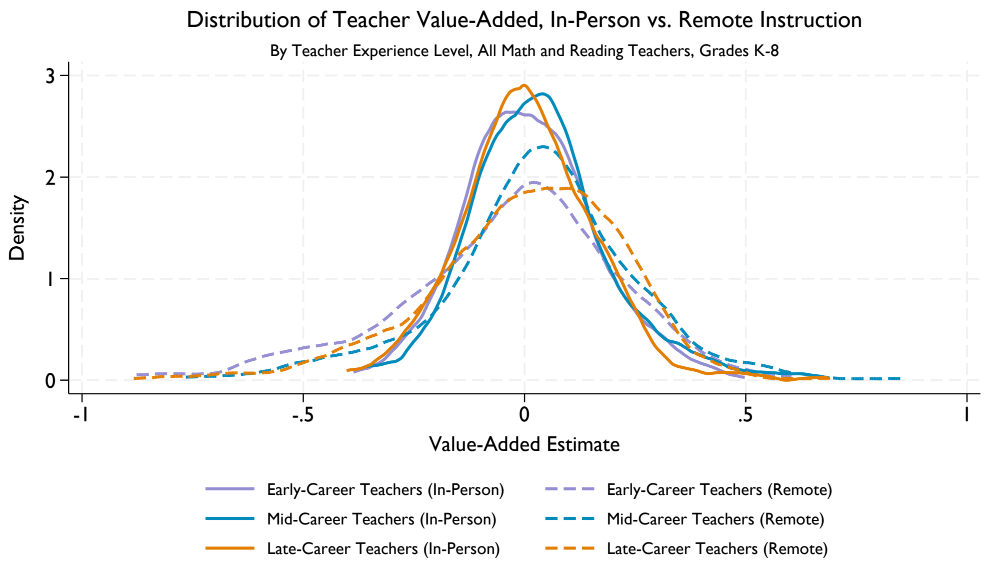
Note: Percentages in each cell above represent the share of teachers initially in the value-added quintile indicated by a given row who ended up in the quintile indicated by each column during the subsequent semester. For example, the value 27.3% in the top left cell of Panel (A) implies that 27.3% of teachers who were in the 1st quintile of effectiveness during Fall 2019 remained in the 1st quintile during Fall 2020. As such, the diagonal entries represent percentages of teachers in each quintile who stayed in the same quintile during the following year. Quintiles used in Table 3 were calculated by first restricting the sample of teachers to those who were present in each of the two years displayed in the relevant panel and then calculating quintiles within subject and district. Each column sums to 100%. Because quintiles are calculated within district and subject, rows sum nearly but not precisely to 100% in cases where the total number of teachers in a district-subject is not divisible by 5.

Figure 1. The distribution of teacher value-added during in-person vs. remote instruction, Districts 1 and 2



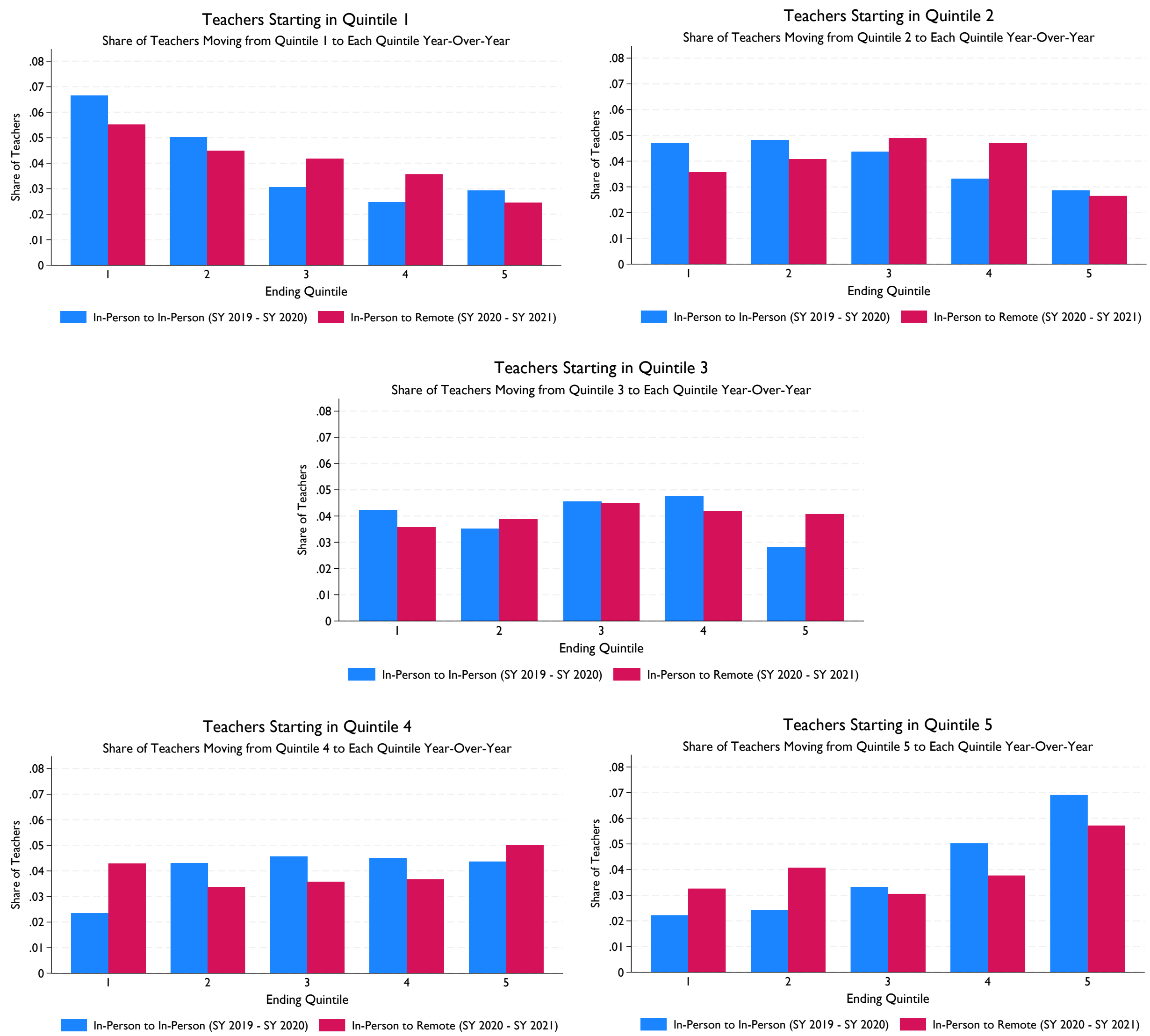
Note: To mitigate the possibility that underlying differences in the distributions are driven by attrition or otherwise differing composition of teachers across years, Figure 1 only includes value-added estimates for teachers observed during both Fall 2019 and Fall 2020.

Figure 2. The distribution of teacher value-added during in-person vs. remote instruction, by teacher experience level, Districts 1 and 2.



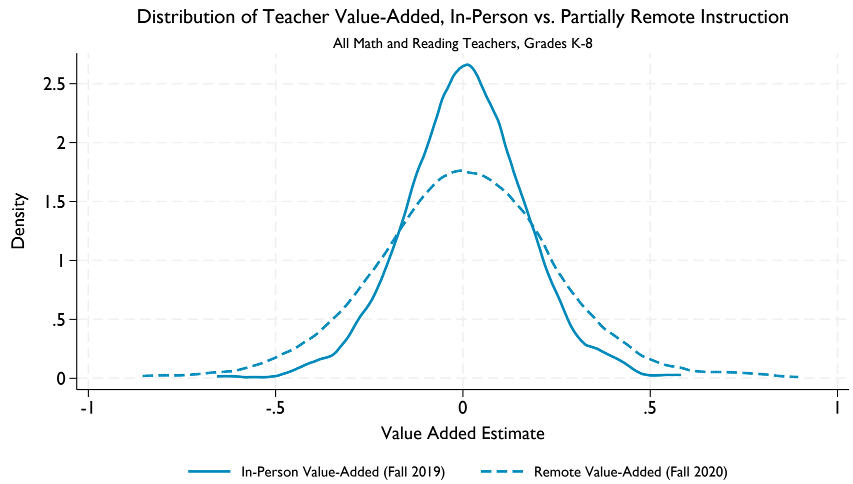
Note: Teacher experience bins are defined as follows: 0-5 years (early-career), 6-15 years (mid-career), > 15 years (late-career). We only display value-added estimates for teachers observed during both semesters.

Figure 3. Percent of teachers making each type of quintile change, comparing an in-person to in-person period with the in-person to remote period, Districts 1 and 2



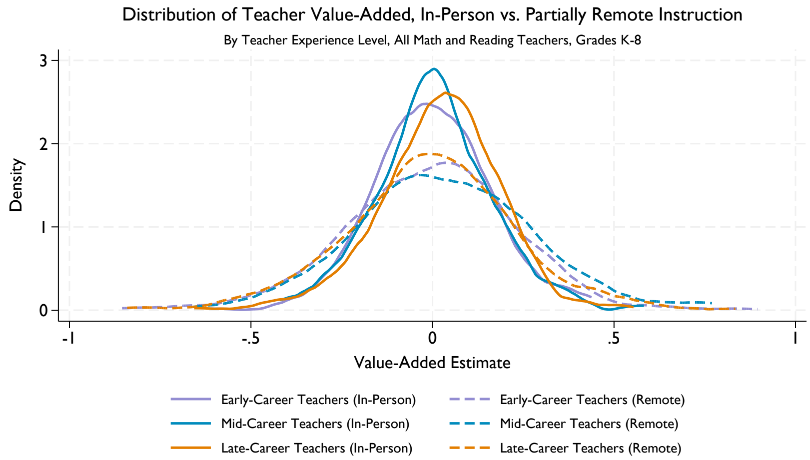
Note: Quintiles reported above are calculated using the fixed set of teachers who are observed in both the initial and final year for a given two-year range.

Figure 4. The distribution of teacher value-added during in-person vs. partially remote instruction, District 3



Note: To mitigate the possibility that underlying differences in the distributions are driven by attrition or otherwise differing composition of teachers across years, Figure 4 only includes value-added estimates for teachers observed during both Fall 2019 and Fall 2020.

Figure 5. The distribution of teacher value-added during in-person vs. partially remote instruction by teacher experience level, District 3



Note: Teacher experience bins are defined as follows: 0-5 years (early-career), 6-15 years (mid-career), > 15 years (late-career). We only display value-added estimates for teachers observed during both semesters.

Figure 6. The distribution of teacher value-added during fully-remote (Fall 2020) vs. partially remote (Spring 2021) instruction, District 2

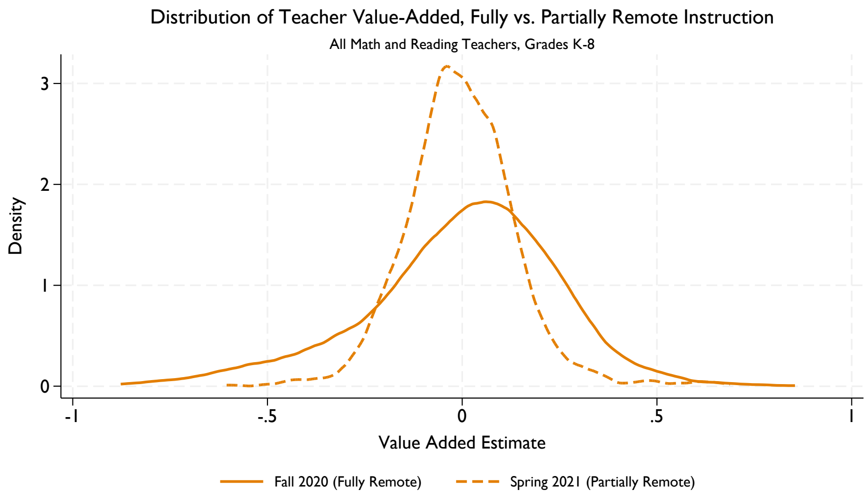
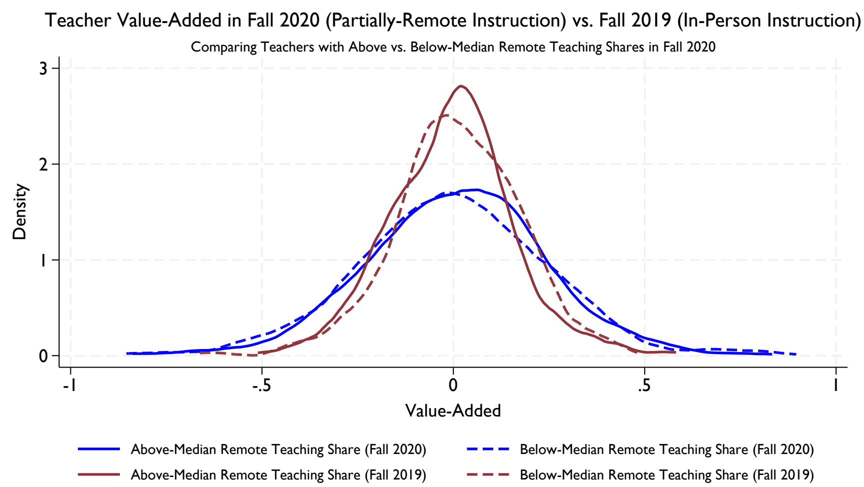


Figure 7. The distribution of teacher value-added during partially-remote vs. in-person instruction in District 3 by share of remote instruction conducted



Appendix

Table A1. Annual teacher value-added correlations, Fall 2016 – Fall 2020

		Fall 2016	Fall 2017	Fall 2018	Fall 2019
In-Person Instruction	Fall 2016	1.000	--	--	--
	Fall 2017	0.3446	1.000	--	--
	Fall 2018	0.2901	0.3674	1.000	--
	Fall 2019	0.2702	0.3594	0.3696	1.000
Remote Instruction	Fall 2020	0.1396	0.2885	0.1901	0.1506

Correlations above are calculated only for teachers in Districts 1 and 2 observed in all of the five semesters (n=304).

Table A2. Comparison of value-added variances by teaching modality and district

	Districts 1/2	District 3
In-Person (Fall 2019)	0.024	0.027
Remote/Partially-Remote (Fall 2020)	0.053	0.058
P-Value	0.000	0.000

The table above displays the variance in value-added for all K-8 teachers in the district indicated by each column header during the time period indicated in each row. P-values represent Levene's test statistic for equality of variances, which is robust to non-normality. Districts 1 and 2 were fully remote during Fall 2020, while District 3 was partially remote.

Table A3. Comparison of variances by teacher subject and teaching modality, Districts 1 & 2

	Reading	Math	P-Value
Fall 2019 (In-Person)	0.025	0.022	0.542
Fall 2020 (Remote)	0.056	0.049	0.100
P-Value	0.000	0.000	

The table above displays the variance in value-added for teachers of each subject (indicated by the column headers) during the time period indicated in each row. P-values represent Levene's test statistic for equality of variances, which is robust to non-normality. Tests of equality of variance reported at the end of each row compare distributions by teacher subject area within-years, while those at the end of each column compare distributions of the same teacher type across years.

Table A4. Comparison of value-added variances by teacher subject or experience level, in-person vs. partially-remote instruction (District 3)

	Early Career	Mid-Career	Late-Career	P-Value	Reading	Math	P-Value
In-Person	0.027	0.026	0.027	0.718	0.026	0.027	0.313
Partially-Remote	0.056	0.061	0.055	0.394	0.048	0.067	0.004
P-Value	0.000	0.000	0.000		0.000	0.000	

The table above displays the variance in value-added for teachers of the type indicated by each column header during the time period indicated in each row. P-values represent Levene's test statistic for equality of variances, which is robust to non-normality. Tests of equality of variance after vertical dashed lines compare variances across categories to the immediate left of the line, while those at the end of each column compare distributions of the same teacher type across years.

Table A5. Comparison of variances, Fully Remote (Fall 2020) vs. Partially Remote (Spring 2021) instruction in District 2

	Variance
Fully-Remote (Fall 2020)	0.062
Partially-Remote (Spring 2021)	0.020
P-Value	0.000

The table above displays the variance in value-added for teachers in District 2 during Fall 2020 and Spring 2021. P-values represent Levene's test statistic for equality of variances, which is robust to non-normality.

Table A6. Comparison of variances, teachers with above vs. below-median remote teaching shares (District 3)

	Below-Median	Above-Median	P-Value
Fall 2019 (In-Person)	0.027	0.026	0.521
Fall 2020 (Partially-Remote)	0.062	0.055	0.328
P-Value	0.000	0.000	

The table above displays the variance in value-added for teachers of the type indicated by each column header during the time period indicated in each row. P-values represent Levene's test statistic for equality of variances, which is robust to non-normality. Tests of equality of variance reported at the end of each row compare distributions by teacher category within-years, while those at the end of each column compare distributions of the same teacher type across years.

Figure A1, Panel A. The distribution of value-added during in-person and remote instruction, math teachers (Districts 1 & 2)

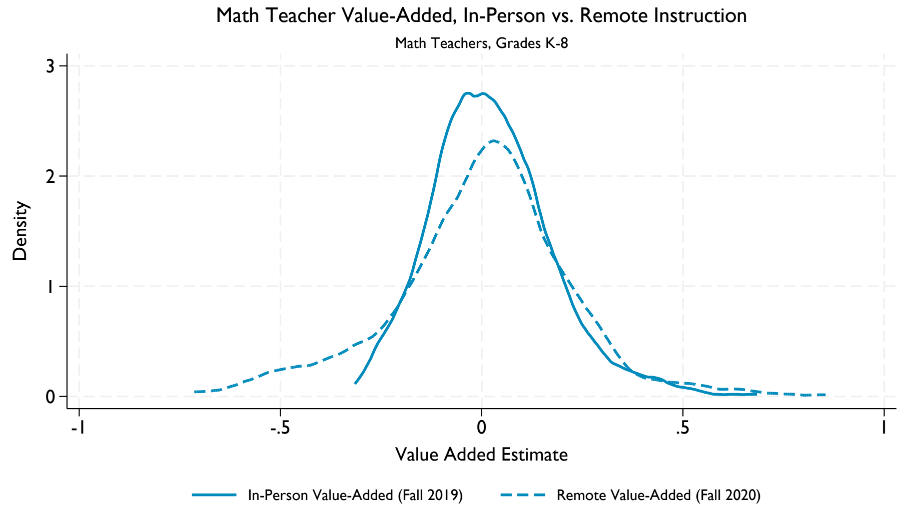
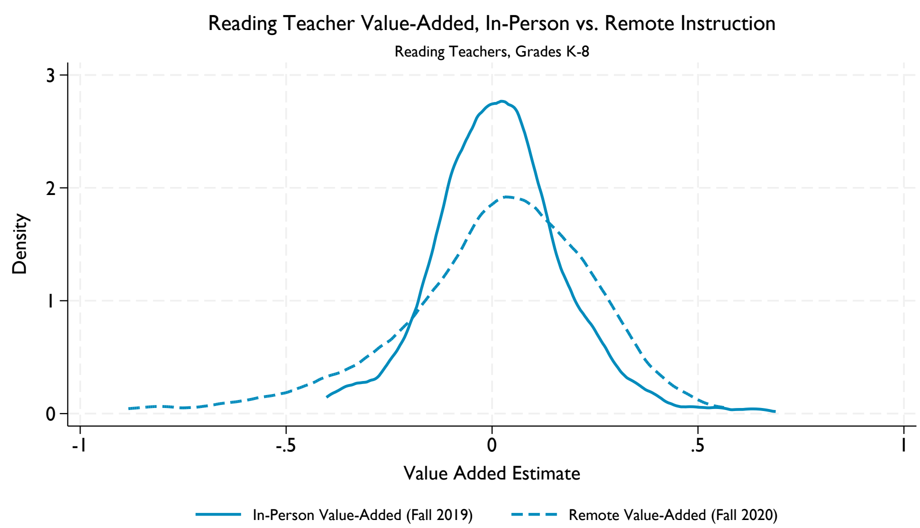


Figure A1, Panel B. The distribution of value-added during in-person and remote instruction, reading teachers (Districts 1 & 2)



To mitigate the possibility that underlying differences in the distributions are driven by attrition or otherwise differing composition of teachers year-to-year, we include value-added estimates for teachers observed during both Fall 2019 and Fall 2020.

Figure A2, Panel A. The distribution of value-added during in-person and remote instruction, elementary (grades K-5) teachers (Districts 1 & 2)

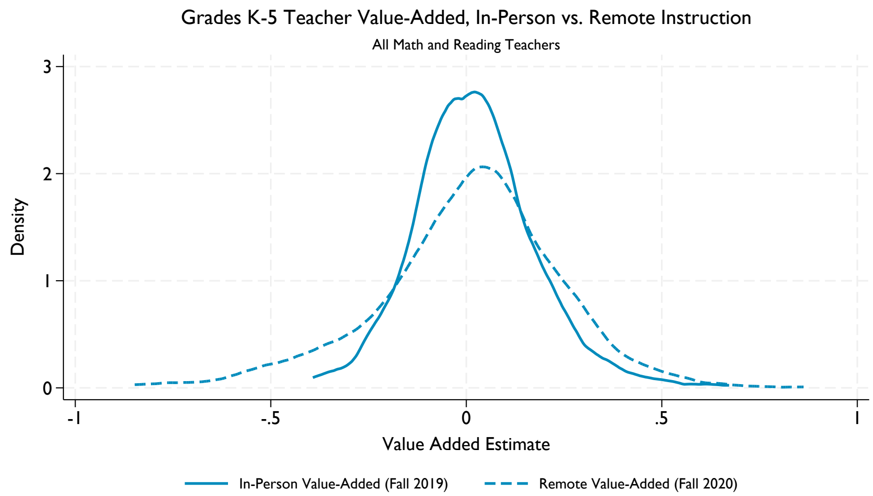
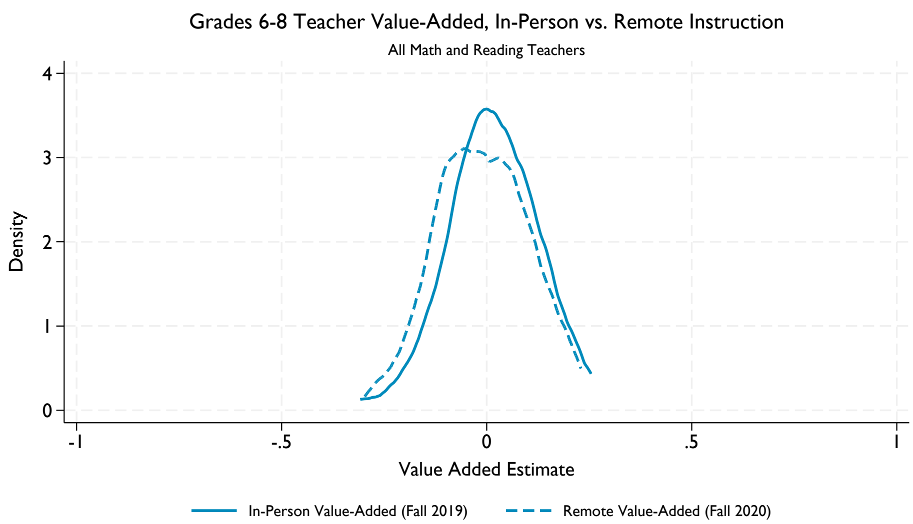


Figure A2, Panel B. The distribution of value-added during in-person and remote instruction, middle school (grades 6-8) teachers (Districts 1 & 2)



To mitigate the possibility that underlying differences in the distributions are driven by attrition or otherwise differing composition of teachers from year-to-year, we include value-added estimates for teachers observed during both Fall 2019 and Fall 2020.

Figure A3. The distribution of teacher remote instruction shares in District 3 during Fall 2020

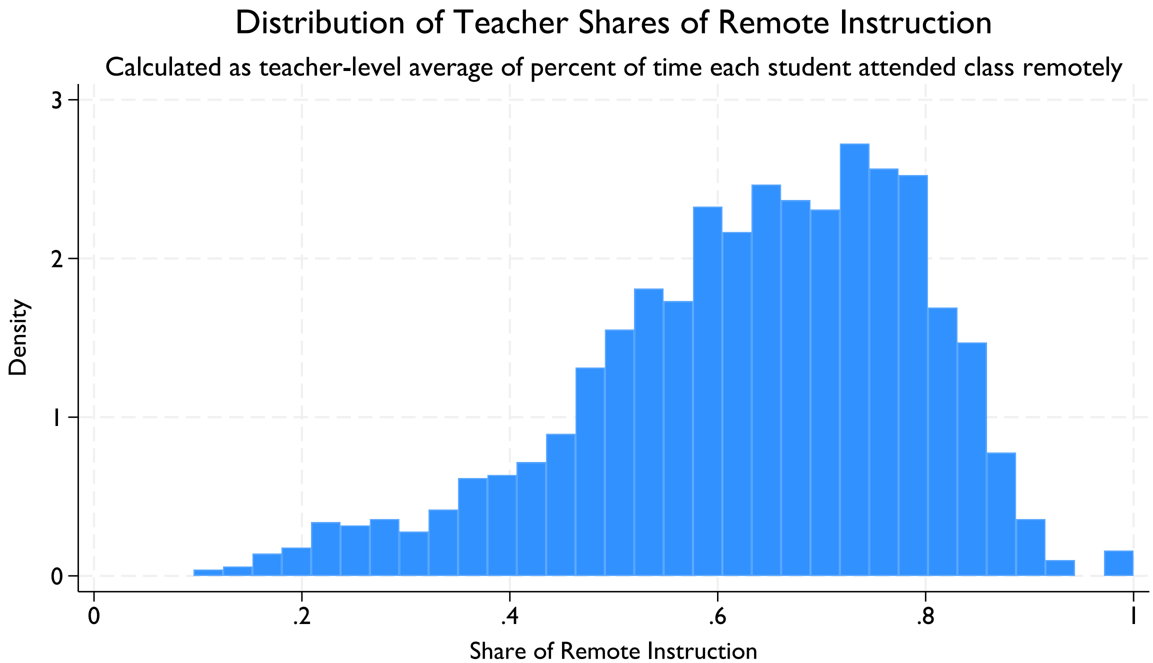


Figure A4. Percent of teachers making each type of quintile change, comparing an in-person to in-person period with the in-person to remote period, District 3