


Causal Inference and COVID:
Contrasting Methods for Evaluating Pandemic Impacts Using State Assessments

Benjamin R. Shear

School of Education, University of Colorado Boulder

Author Note

Benjamin R. Shear  <https://orcid.org/0000-0002-9236-2927>

The author acknowledges helpful feedback on drafts of this manuscript from Lorrie Shepard, Derek Briggs, and Mimi Engel at the University of Colorado Boulder, and from Marie Huchton at the Colorado Department of Education.

Correspondence concerning this article should be addressed to Benjamin R. Shear, School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309.

Email: benjamin.shear@colorado.edu

Abstract

In the spring of 2021, just one year after schools were forced to close for COVID-19, state assessments were administered at great expense to provide data about impacts of the pandemic on student learning and to help target resources where they were most needed. Using state assessment data from Colorado, this paper describes the biggest threats to making valid inferences about student learning to study pandemic impacts using state assessment data: measurement artifacts affecting the comparability of scores, secular trends, and changes in the tested population. The paper compares three statistical approaches (the Fair Trend, baseline student growth percentiles, and multiple regression with demographic covariates) that can support more valid inferences about student learning during the pandemic and in other scenarios in which the tested population changes over time. All three approaches lead to similar inferences about statewide student performance but can lead to very different inferences about student subgroups. Results show that controlling statistically for pre-pandemic demographic differences can reverse the conclusions about groups most affected by the pandemic and decisions about prioritizing resources.

Keywords: causal inference; state assessments; COVID-19 effects

Causal Inference and COVID:

Contrasting Methods for Evaluating Pandemic Impacts Using State Assessments

After a pause in state testing during the 2019-20 academic year due to the pandemic, the U. S. Department of Education (DOE) required all states to conduct statewide testing in the spring of 2021. There were two primary motivations cited for testing: 1) document the impact of the pandemic on student learning and 2) target resources to schools and communities most in need. Guidance from the DOE argued that state assessments were needed "to understand the impact COVID-19 has had on learning and identify what resources and supports students need," and to "be prepared to address the educational inequities that have been exacerbated by the pandemic, including by using student learning data to enable states, school districts, and schools to target resources and supports to the students with the greatest needs" (U.S. Department of Education, 2021, p. 1). Several prominent advocacy and civil rights organizations made similar arguments in favor of state testing (National Urban League et al., 2020).

The first aim is explicitly about understanding the causal effect of the pandemic on student learning. For the second aim, if the goal is to allocate resources to schools and students with the greatest needs, this could be done without any state assessment data, by using opportunity to learn (OTL) metrics, disparities in existing resources, or disparities in the pandemic's economic, social, and health impacts on students' communities. Using assessment data to target resources does not require causal inferences (unless the goal is to target resources where student learning was *most affected* by the pandemic), but it does require making inferences about student learning during the pandemic across different schools or groups.

Cross-sectional comparisons of aggregate state assessment results across years are regularly used to make inferences about trends in student achievement for system monitoring

purposes. Drawing causal inferences based on these comparisons is less common. Both types of inferences are more complicated than in previous years. Changes to test administration processes and consistently lower test participation rates in 2021 (Gewertz, 2021) make it difficult to disentangle the extent to which changes in scores, statewide or for individual schools or student subgroups, are due to actual changes in student learning versus changes to test administration conditions or to the population of tested students. Moreover, with no state test score data from 2020, states using growth models based on prior year data could not report standard growth metrics that provide more complete descriptions of student learning as part of accountability systems.

Testing experts and research partners working with state departments of education have developed additional reporting metrics to provide more complete descriptions of 2021 results that are responsive to these challenges, but there has been little consistency or synthesis of methods and results across states. Examples include metrics adjusting for prior scores (Betebenner & Wenning, 2021; Ho, 2021) and more general multiple regression analysis (Kogan & Lavertu, 2022). Although some of these approaches were developed to facilitate descriptive inferences about 2021 results, they are also being used to make causal inferences about the impact of the pandemic on student learning and to compare estimated effects across student subgroups (e.g., see studies included in the review by Cohodes et al., 2022).

This paper compares three statistical adjustments that can be used to facilitate more valid cross-year comparisons when tested populations change over time. The methods are illustrated using spring 2021 state assessment results but are applicable in other years and for other types of assessments, such as nationally used interim assessments. Many widely reported studies quantifying the impact of the pandemic on student learning, for example, were based on interim

assessments (e.g., Curriculum Associates, 2021; Lewis et al., 2021; West et al., 2021). These studies provided important early evidence about the effects of the pandemic on learning, particularly during 2020 when state tests were not administered. However, interim assessment data faced similar limitations to those for state assessments. Interim assessment results only represent students enrolled in schools that administered a particular interim test, for example, and the population of students participating in these assessments changed during the pandemic as did test administration conditions.

The next two sections clarify the definition of the causal effect of the pandemic on student learning and discuss some of the biggest challenges to interpreting cross-sectional comparisons of state assessment results over time. The following section compares three statistical approaches used to provide more detailed descriptions of student achievement trends in 2021 and to estimate the effects of the pandemic. The comparison demonstrates how these approaches are statistically and conceptually similar but also highlights an important case in which they can lead to substantially different conclusions, i.e., when comparing student learning across different subgroups. These methods are illustrated by applying them to state assessment data from Colorado, where participation rates were variable due to student enrollment in remote learning and other factors.

Defining the Causal Effect of the Pandemic

Evaluating estimates of the effect of the pandemic requires being clear about the definition of the effect. Policy researchers recommended focusing on the “overall” effect of the pandemic on student learning, representing the impacts due to all economic, social, health, and educational experiences (Bacher-Hicks & Goodman, 2021). Using the Neyman-Rubin causal model (Holland, 1986; Imbens & Rubin, 2015) to define the causal effect of the pandemic there

are two potential outcomes for each student in this context: 1) how each student *actually scored* on the end-of-year state test in 2021 after experiencing more than a year of pandemic-disrupted schooling, and 2) how each student *would have scored*, had there not been a pandemic. Only the first score is observed, while the latter is referred to as the counterfactual outcome. The overall causal effect of the pandemic on each student's test score is the difference in these two scores, while the average effect is the average difference in these two scores for all students in the population of interest. This framework is a reminder that estimating the effect of the pandemic requires being clear about the population of interest and the definition of the effect implied by choice of an estimate for the counterfactual outcome.

Challenges to Estimating the Effect of the Pandemic

Researchers typically rely on a comparison group to estimate counterfactual outcomes. For the pandemic, however, there is no concurrent comparison group, given its global nature and universal disruptions. Further, there are no state test score data from 2020 for comparison. Historical data from state testing in 2019 and earlier provide the only available data to make comparisons. Although intuitive, direct comparisons of 2021 to 2019 grade-level results have the potential to provide misleading conclusions about changes in student learning during the pandemic and hence about the effect of the pandemic on student learning. These estimates face three primary threats to internal validity: measurement artifacts affecting the comparability of scores, secular trends, and changes in the tested population.

Comparability of Scores. Cross-year comparisons assume score scales are directly comparable across years. Many of the psychometric methods routinely used to make scores comparable across years rely on assumptions that may not hold true given the pause on testing in 2020, changes to academic and instructional calendars, the timing of testing in 2021, changes to

test administration procedures in 2021, and other factors (Balow & Miller, 2020). This could result in differences in average grade-level scores due to measurement artifacts rather than due to changes in student learning. A detailed consideration of these issues is beyond the scope of this paper, but any comparison of 2021 test results to prior years should begin with psychometric checks by testing contractors and state assessment staff to ensure that individual student scores in 2021 are comparable to individual student scores in prior years.

Secular Trends. In many states, average test scores within grades increase slightly each year across cohorts. Trends can reflect improving instructional practices, increasing alignment between instruction and assessments, or even “score inflation” (Koretz, 2008) due to factors such as teaching to the test. Although changes in average test scores from year to year are often small, their presence suggests that we could reasonably expect the 2021 cohort in any given grade to have had higher average scores than the 2019 cohort in the absence of major changes. As a result, comparing 2019 average scores to 2021 average scores could potentially underestimate any negative effects of the pandemic (or overestimate if secular trends were negative) even if test participation rates remained high.

Changing Populations. Perhaps the greatest impediment to drawing valid conclusions about changes in student learning across years is selection bias due to changes in the population of students taking tests. Historically, the statewide population of students may not change noticeably year to year when nearly all students participate in state testing, making the average test scores in recent prior years a reasonable counterfactual or comparison for the current year. In 2021 there were two phenomena that threaten the validity of this assumption: declining public school enrollment (Pendharkar, 2022) and reduced test participation rates (Gewertz, 2021). These correspond to concerns related to selection and attrition, respectively. As a result, changes

in average test scores in 2021 may be due to differences in the population of students participating in testing, not solely the impacts of the pandemic.

Threats to Descriptive Inferences Based on Cross-Sectional Comparisons. Cross-sectional comparisons of average scores could be used to make descriptive, rather than causal, inferences about trends in student achievement. Descriptive interpretations remain agnostic about the specific events or policies that may have caused changes in achievement. Nonetheless, inferences that changes in average scores reflect changes in student learning are also threatened by a lack of score comparability or changes in the tested population of students. Secular trends pose less of a threat to descriptive inferences, although they may affect how the changes are interpreted.

Review of Methods

This section describes three approaches being used to support more valid interpretations of 2021 state assessment results by adjusting for changes to the population of tested students: the “Fair Trend” (FT) approach (Ho, 2021), the “baseline” student growth percentile (SGP) model (Betebenner & Wenning, 2021), and a more general multiple regression (MR) model that includes prior test scores and student demographic variables. Each method produces an adjusted metric that characterizes test scores of 2021 students relative to different pre-pandemic norms. Although the approaches can be used for purely descriptive purposes, each one has also been used to make causal inferences about the impact of the pandemic for statewide populations of students as well as demographic subgroups.

The adjustments are designed to account for secular trends and changing populations but cannot adjust for changes to tests, as all three methods assume grade level and prior test scores used in the analyses are comparable across the 2019 and 2021 cohorts. These methods rely on

students' prior test scores and are applicable only for tested grades in which students have two-year prior test scores available for matching; in most states this will be in grade 5 or higher, given that grade 3 is the earliest required tested grade.

The Fair Trend (FT). Ho (2021) recommends calculating the FT metric instead of directly comparing average scores in 2021 to prior years. This metric was developed to support “appropriate comparisons of performance this year to the performance of academic peers two years prior” (p. 4). In this context “academic peers” are students with similar prior grade test scores. For each tested grade and subject, student longitudinal test score data from 2017 and 2019 and a method such as ordinary least squares (OLS) are used to estimate the model

$$y_{ig,2019} = \alpha_0 + f(x_{i(g-2),2017}) + e_{ig,2019}. \quad \text{Eq. 1}$$

Here $y_{ig,2019}$ is the score for student i in grade g in 2019, $x_{i(g-2),2017}$ is the same student's score from two grades and two years prior, and $f(\)$ is a function of prior scores. Then, for each student in grade g in 2021 who participated in testing and has a prior score from grade $g - 2$ in 2019, parameter estimates from Eq. 1 are used to compute a predicted score as

$$\hat{y}_{ig,2021} = \hat{\alpha}_0 + \hat{f}(x_{i(g-2),2019}). \quad \text{Eq. 2}$$

This is the score that we would expect a student in 2021 to earn if they made the same progress as their academic peers from 2019. The FT statistic for an aggregate group is the average difference between predicted and observed scores among students in the 2021 cohort:

$$FT_g = \text{mean}(y_{ig,2021} - \hat{y}_{ig,2021}). \quad \text{Eq. 3}$$

The FT statistic characterizes 2021 test scores for tested students relative to peers who share the same prior test scores. More flexible functional forms or estimation algorithms could be used for the prediction equation. Keng et al. (2022) provided additional details and used the FT approach to estimate pandemic impacts in Utah.

Baseline Student Growth Percentiles (SGPs). Many states annually report SGPs (Betebenner, 2009) to summarize student progress each year. The SGP model uses quantile regression to compare a student’s current grade test score to the scores of their academic peers (defined the same way as they are in the FT model) and assign a percentile rank based on the quantile regression surfaces. An SGP of 50, for example, represents a student who scored at the median relative to other students with similar achievement in prior grades. The median or mean SGP (MGP) is used to summarize SGPs for a group of students, where an MGP 50 indicates “typical” performance for a group relative to students’ academic peers.

The SGP statistics typically reported are “cohort-referenced” SGPs rather than “baseline-referenced” SGPs (Betebenner & Wenning, 2021; Shear, 2020). The cohort-referenced SGP model compares students’ current grade scores to their academic peers in the same cohort. In the context of the pandemic, however, the cohort SGP model will “adjust away” any overall effect of the pandemic, because the growth students make would be compared to other students within their own cohort who were also affected by the pandemic. The baseline SGP model overcomes this limitation by comparing student performance in 2021 relative to their academic peers in a prior “baseline” cohort of students who did not experience the pandemic.¹ Baseline MGP results for particular groups (or a statewide population) can be used as indicators of the impact of the pandemic on student learning, with values below 50 quantifying the extent to which student scores in 2021 were lower than the scores of their academic peers in pre-pandemic cohorts (Betebenner & Wenning, 2021). Betebenner and Van Iwaarden (2022) provided additional details and used the baseline SGP model to estimate pandemic impacts in Rhode Island.

¹ The pause on state testing in 2020 also required the transition to a “skip year” growth model that conditions scores on two-year prior scores rather than one-year prior scores normally used. However, this difference is irrelevant to the cohort/baseline SGP distinction highlighted here.

Multiple Regression (MR). A more general MR model that includes prior test scores and student demographic variables is a third approach to characterize student test scores in 2021. For each tested grade and subject, this proceeds by estimating a MR model describing pre-pandemic test score trends for the 2019 cohort that includes prior test scores and student demographic variables:

$$y_{ig,2019} = \beta_0 + g(x_{i(g-2),2017}) + \delta \mathbf{Z}_{ig,2017} + e_{ig,2019}. \quad \text{Eq. 4}$$

Here, $g(\)$ is a function of prior test scores and the vector $\mathbf{Z}_{ig,2017}$ includes student demographic variables as recorded in 2017 (for students in the 2021 cohort they are based on 2019 data so that they represent pre-pandemic characteristics). The estimated model is then used to calculate predicted scores for each tested student in the 2021 cohort as:

$$\tilde{y}_{ig,2021} = \hat{\beta}_0 + \hat{g}(x_{i(g-2),2019}) + \hat{\delta} \mathbf{Z}_{ig,2019}. \quad \text{Eq. 5}$$

The average difference between observed and predicted scores,

$$MR_g = \text{mean}(y_{ig,2021} - \tilde{y}_{ig,2021}), \quad \text{Eq. 6}$$

characterizes 2021 test scores for tested students relative to peers who share the same prior test scores and demographic characteristics. This approach was used in a recent analysis of national interim test score data (Goldhaber et al., 2022) and a similar model was used by Kogan and Lavertu (2022) to estimate pandemic impacts in Ohio. The Appendix describes connections among these uses of MR.

Comparison of Interpretations. All three approaches share a conceptual similarity whereby they compare students in 2021 to a matched set of “academic peers” in 2019. Mechanically, all three methods make this comparison by estimating a prediction model based on the 2019 pre-pandemic cohort of students, and then describing the scores of students in the 2021 cohort relative to these predictions. From a statistical perspective, all three models are

“conditional status models” (Castellano & Ho, 2013). The methods differ in terms of whether the comparison is reported in scale score units (FT and MR) or percentile rank units (SGP), whether the predictions are based on linear regression (FT and MR) or quantile regression (SGP), and whether the predictions include only prior test scores (FT and SGP) or include prior test scores and demographic covariates (MR).

The latter difference about variables used to make the predictions is the most relevant. Although the FT and SGP will produce apparently different results due to using scale scores versus percentile ranks, both approaches describe 2021 student performance relative to pre-pandemic academic norms and are likely to be highly correlated. The MR approach describes 2021 student performance relative to pre-pandemic *sociodemographic* academic norms and is thus making a fundamentally different comparison. In the analyses below, all three methods produce similar inferences about average performance statewide but can produce very different inferences about demographic subgroups.

When these adjustments are interpreted as causal effect estimates, they are most accurately characterized as estimating the causal effect of all experiences *during* the 2019 to 2021 period relative to the 2017 to 2019 period, rather than the effect *of* the pandemic specifically. These adjustments rely on strong assumptions about unconfoundedness to produce unbiased causal effect estimates even for this definition of the effect. However, these statistics do not need to be interpreted as causal effects. Descriptive characterizations of student test scores relative to a matched set of pre-pandemic academic peers provide useful information about student academic progress during the pandemic.

A Note about Matching. Although beyond the scope of this paper, statistical matching approaches such as propensity score matching (PSM; Rosenbaum & Rubin, 1983) or coarsened

exact matching (CEM; Iacus et al., 2012) could also be used to estimate the effect of the pandemic. These methods are not considered here because the focus is contrasting aggregate metrics that produce descriptive and causal characterizations of 2021 test score results, rather than all possible approaches to producing causal estimates.

Participation Rate and Match Rate

The three methods described above can only be implemented using longitudinal data for students who participated in testing in 2021 and have prior test scores from 2019, thus raising concerns about external validity. Results from these methods should be accompanied by metrics that help quantify the extent to which students included in the analysis represent the population of all students, such as test participation and match rates (An et al., 2022; Ho, 2021). Ho recommended calculating the match rate as the proportion of students tested in grade $g - 2$ in 2019 who also tested in grade g in 2021. This will be referred to as the “prior match rate.” As an alternative, one can calculate a “current match rate” as the proportion of enrolled students in grade g in 2021 who had both current grade g test scores and grade $g - 2$ test scores from 2019. Both match rates share the same numerator (“matched” students) but differ in terms of the denominator.

The test participation rate is the simplest of the three rates to explain: it tells us about missing data due to attrition by reporting the proportion of enrolled students who participate in testing. The prior match rate reflects both selection and attrition - students who are missing either because they were not tested or because they left the school system between 2019 and 2021. The current match rate reflects attrition due to non-participation and missing prior scores and is useful for evaluating the generalizability of estimated effects among enrolled students because it quantifies the proportion of enrolled students represented by estimates using matched scores.

Colorado Data

These methods are illustrated by applying them to 2021 math and English Language Arts (ELA) test scores for elementary and middle school students in Colorado. These examples use statewide longitudinal student-level test score and demographic data provided by the Colorado Department of Education (CDE). Colorado received a waiver from the U. S. Department of Education to reduce the amount of testing in 2021 by alternating subjects in every other grade. Testing in ELA was required only in grades 3, 5, and 7, while testing in math was required only in grades 4, 6, and 8. Results for the Colorado Measures of Academic Success (CMAS) tests are only reported for grades and subjects tested in both 2019 and 2021.² The methods described above are only applied to data from grades 5-8, as tested students in 2021 in 3rd and 4th grade did not have prior CMAS test scores available from 2019. CDE determined that CMAS test scores for individual students in 2021 remained comparable to scores from prior years (Colorado Department of Education, nd). All analyses were carried out in the R computing environment (R Core Team, 2022).

Table 1 reports statewide grade level sample sizes, test participation rates, average test scores, and standard deviations (SD) for 2019 and 2021. As observed in many other states, average test scores were lower in 2021 relative to 2019 at each grade level, although the magnitude varied from -5.9 scale score points (6th grade math) to -0.8 scale score points (5th grade ELA). Table 1 also shows variability in test participation rates. In 2021, participation rates were about 20 percentage points lower in elementary grades and as much as 30 percentage points

² About 1-3% of 3rd and 4th grade Colorado students each year take Spanish language assessments in place of ELA tests. These scores are not included in these analyses.

lower in middle school.³ The large declines in participation rates in 2021 call into question inferences that the score declines can be explained by the effects of the pandemic alone. Also of note, participation rates in Colorado increased between 2015 and 2019 but remained below 90% in some grades. This may limit the generalizability of results to the extent that the pre-pandemic norms do not represent the full population.

Cross-year comparisons of average scores should be accompanied by descriptive statistics that characterize the relevant populations. Table 2 describes the characteristics of relevant populations and subpopulations of students at the 5th grade level. The first three columns of Table 2 describe enrolled students in 2021 and all tested students in 2021 and 2019, respectively. The final two columns describe the subpopulations of tested students in 2019 and 2021 who also had test scores from two years prior; these are the students included in the adjusted analyses presented in the next section and are referred to as the “matched samples.” The final three columns of Table 2 report standardized average prior test scores by demographic subgroup. Comparing the tested and enrolled populations in 2021 helps to understand whether the 2021 test score results can generalize to the population of all enrolled students. Comparing characteristics of tested students in 2019 and 2021 helps to understand systematic differences that could be confounding the direct comparison of observed average scores across years. Similar comparisons can be made between the matched samples to aid interpretation of the statistical adjustments reported below.

Table 2 shows systematic differences between populations that threaten the validity of descriptive and causal inferences based on directly comparing average scores in 2019 and 2021.

³ These participation rates differ slightly from officially reported participation rates from CDE because the calculations here are based on all students included in the administrative data files, while the figures reported by the state consider additional business rules used for calculating and reporting participation.

Relative to tested students in 2019, tested students in 2021 had higher prior (3rd grade) scores and were less likely to be from historically underserved groups, such as students of color and students living in poverty (as indicated by FRL status). The tested students in 2021 also do not appear to be a representative sample of the population of enrolled students. In addition to differences in average prior scores, White students are overrepresented among the population of tested students, while students from all other demographic groups listed in Table 2 tend to be underrepresented. Similar differences exist among the matched samples of students and across grades 5-8. The final three columns indicate variability in the extent to which matched 2021 students within subgroups are: 1) representative of enrolled students with prior scores, and 2) systematically different from matched students in 2019. These differences highlight the importance of using statistical adjustments to facilitate more valid inferences about trends and caution generalizing results to the population of enrolled students.

Results

Rows 1 and 2 in Table 3 present, respectively, the prior and current match rates for each grade and test subject. The prior match rates range from 55.1%–68.3% and are about 2–5 percentage points lower at each grade level than participation rates in Table 1, suggesting that there are additional students missing from 2019 beyond those who did not participate in testing. For reference, 2019 prior match rates ranged from about 85%–90%. The current match rates are an additional 2–4 percentage points lower than the prior match rates in each grade and suggest caution generalizing from the results of analyses requiring prior test scores to the population of all enrolled students.

The third row of Table 3 repeats the observed differences between 2019 and 2021 scores that were shown in Table 1. The fourth row reports the (unadjusted) difference in average scores

between matched samples. These serve as a point of reference to be compared to the next three rows, which present estimates from the FT model, MR model, and the mean baseline SGP (MGP) subtracted relative to 50.⁴ The adjusted differences are larger in absolute value than the observed (unadjusted) differences in average grade level scores. Framed as a causal inference, the unadjusted differences underestimate the magnitude of pandemic effects (i.e., overestimate student progress) due to a combination of factors including population changes and changes in the relationship between prior and current scores. The MR and FT estimates are very similar. The MGP estimates cannot be directly compared to the FT or MR estimates because they are in percentile rank units rather than scale scores, but the relative magnitude of the estimates across grades and subjects is the same.

Because scale score estimates are not directly comparable across grades or with results from different state tests, the next four rows of Table 3 report the same differences in SD or “effect size” units, calculated by dividing each difference by the 2019 grade level SD. The magnitude of the MR estimates in SD units were larger in math (-0.21 and -0.15) than in ELA (-0.08 and -0.13) consistent with findings reported in early reviews of evidence about pandemic impacts (e.g., Betebenner et al., 2021; West et al., 2021). The estimates are slightly smaller than those reported in an analysis of state data in Ohio using MR, which ranged from -0.28 to -0.35 SDs in math and -0.08 to -0.15 SDs in ELA (Kogan & Lavertu, 2022), and are similar to

⁴ The SGP statistics are those officially calculated for CDE by the Center for Assessment, with the exception that the mean rather than median is reported here for greater comparability with the other metrics. These baseline SGPs are relative to the 2019 cohorts and use all available skip-year prior scores (they include all prior scores from two or more years before testing and in grades 6+ they can include multiple prior scores if students have more than one available). The FT and MR models used a cubic polynomial function of prior scores to predict current grade scores. The MR model uses the demographic variables reported in Table 2: gender (male or female), race/ethnicity (American Indian/Alaska Native, Asian, Black, Hawaiian/Pacific Islander, Hispanic, White, or Two or More Races), free/reduced-price lunch eligibility (FRL) status (yes/no), individualized education plan (IEP) status (yes/no), and English language learner (ELL) designation (yes/no). The FRL, IEP, and ELL variables are from the year of prior scores (i.e., from 2019 for students in the spring 2021 cohort).

estimates based on applying the FT model in Utah, which were -0.09 to -0.16 in ELA and -0.19 in math (Keng et al., 2022).

For states that have vertically scaled tests, the differences could be converted to “months of learning” based on pre-pandemic average changes in scores across grades. In states such as Colorado that do not have vertically scaled tests, pre-existing normative benchmarks (e.g., Hill et al., 2008) could be used to convert effect sizes to months of learning, but these conversions rely on several assumptions and should be done cautiously if at all (Student, 2022). The differences could also be compared to the distance between relevant cut scores (Keng et al., 2022). The distance from the “Approaching” to “Meeting” grade level cut score is 25 points on the CMAS scale in grades 3–8. The MR estimates represent between 10% and 27% of this distance depending on grade level.

Despite well-known problems with interpreting percent “proficient” statistics (Ho, 2008; Holland, 2002), some audiences may prefer to report or interpret the differences in percent “proficient” units. Estimating the differences in the percent proficient metric for either the FT or MR model can be done by replacing each student’s observed current grade score, y_{igy} , with a binary indicator $I_{y_{igy}}$ equal to 1 if a student’s current-grade score reached the grade-level “proficiency” standard and equal to 0 otherwise.

Table 3 reports the observed differences in proficiency rates between 2021 and 2019 with FT and MR estimates adjusting for prior scores and demographics.⁵ The observed declines in percent proficient ranged from -1.2 to -7.4 percentage points in 2021. The adjusted estimates highlight two findings. First, as with comparisons based on scale scores, the observed

⁵ Logistic or probit regression could also be used to estimate the effect on proficiency rates in a similar manner. A linear probability model was used for the present analyses, but results were substantively the same using a logistic regression model.

(unadjusted) declines in proficiency rates from 2019 to 2021 are smaller than the adjusted differences and likely underestimate the effects of the pandemic. Second, while the scale score and SGP adjusted differences suggest the largest declines in 6th grade math, the proficiency rate differences suggest the largest declines in 8th grade math, with declines in 6th grade math and 7th grade ELA that are smaller and nearly identical. This underscores the challenge of comparing proficiency rate estimates across grade levels for which the base rates differ, a challenge that will also apply when comparing proficiency rates across states.

It is also essential to consider how missing data could impact inferences. The statistical adjustments described above do not provide direct information about students who did not participate in testing or do not have prior scores. There are multiple reasons the effect of the pandemic could have differed for students who did not participate in spring testing. In Colorado, for example, many students enrolled in remote learning did not participate in testing. Table 2 indicates that missing students had lower prior scores on average and were more likely to be from historically underserved populations. Methods such as multiple imputation could be used to re-estimate the models, but these approaches may be difficult to explain to general audiences and rely on strong assumptions about the association between missing data and observed scores. Ho (2021) recommends an “equity check” metric that reports the academic peer average score for students who did not participate in testing in 2021, but this metric primarily provides information about missing students’ prior scores and does not directly answer questions about how inferences might be sensitive to missing scores.

Table 3 reports ad hoc “bounds” on what the estimated differences might have been if all enrolled students had participated in testing. A lower bound (LB) is calculated by assuming there was no effect on learning for missing students (i.e., they would have made the same progress as

their peers in the 2019 cohort). An upper bound (UB) is calculated by assuming that the average decline was twice as large among missing students.⁶ Although unsubstantiated, these values provide an intuitive sense for how much the estimated differences could be biased by missing data and are reported in the final two rows of Table 3 relative to the MR estimates. The LB values suggest that, even if missing students were not affected by the pandemic, the average statewide declines relative to pre-pandemic cohorts would still range from -0.05 to -0.13 SD across grades, while the UB values warn that average statewide declines could be substantially higher if missing students were more impacted than tested students, ranging from -0.10 to -0.30 SD. These bounds implicitly impute scores that reflect different assumptions. Similar bounds could be used in other years to provide intuition about the sensitivity of aggregate metrics to missing data.

Differential Effects and Targeting Resources

Estimating the effect of the pandemic for different subgroups is another important aim of educational research and decision-making. There has been considerable concern, for example, that the impacts of the pandemic were larger in less affluent communities and communities of color and that additional resources should be provided to these communities. Alternatively, policymakers may decide to target resources to groups or communities where students made the least academic progress without invoking a causal explanation.

Table 4 reports the three adjusted metrics by student race/ethnicity (for the four largest racially identified subgroups in Colorado), student gender, and students' pre-pandemic FRL status, using 5th grade ELA scores as an example. Each row reports the number of students

⁶ To calculate these bounds, let p be the current match rate and let δ be the estimated difference using the MR model. The lower bound can be calculated as $LB = p * \delta$. The upper bound is calculated as $UB = (p * \delta) + [(1 - p) * x * \delta]$, where $x = 2$ in the example for an effect two times as large among missing students.

included in the matched sample, the current match rate, and the adjusted differences. The FT and MR estimates are produced by calculating the mean differences in Eq. 3 and Eq. 6, respectively, for each subgroup (and are again reported in effect size units). The MGP estimates are calculated by subtracting 50 from the observed MGP within each subgroup. Although the MGP estimates cannot be directly compared to the FT or MR estimates, the final three columns of Table 4 report the ratio of the subgroup estimates relative to the statewide estimate for each metric, showing the relative similarity of the FT and MGP estimates.

The first row of the table repeats the estimates from Table 3 for all students in 5th grade as a reference point. The second and third rows clearly illustrate differences between the FT and MR models. The FT estimate for Asian students in Colorado is 0.02 SD (a positive difference), while the MR estimate is -0.12 SD (the largest negative MR estimate in the table). The opposite pattern is observed for Black students in Colorado; while the FT estimate is -0.17 SD, the MR estimate is about half as large at -0.09 SD. The MR estimates are more similar across groups than are the FT or MGP estimates, although we should be cautious about comparing estimates across subgroups or generalizing to the population of students within each subgroup because current match rates range from 51% to 75%. In addition, as subgroup sample sizes become smaller the estimates will be less precise.

These differences are not inconsistent nor are they unique to the pandemic context. They occur because the FT and SGP models describe student performance in 2021 relative to different pre-pandemic norms than the MR model. The FT and SGP models compare students' scores to the scores of academic peers statewide in 2019, whereas the MR model compares students' scores to the scores of academic peers within the same demographic subgroup in 2019. Clearly, however, when these results are used to make causal inferences about the effect of the pandemic,

or determine how much progress each group made, they can lead to very different conclusions. The FT and SGP models estimate effects for subgroups that implicitly combine impacts of the pandemic on student learning with subgroup differences in pre-pandemic growth rates. The MR model is intended to isolate the unique effect of the pandemic by adjusting for these pre-pandemic differences. Relative to MR, the FT or SGP approaches will suggest more progress during the pandemic for subgroups that made greater pre-pandemic progress. Differences can also occur when making comparisons across schools or districts, or student subgroups within schools or districts, although the exact patterns of differences will depend on the circumstances in each state.

Discussion

This paper contrasted three statistical adjustments that can be reported to facilitate more valid inferences about trends in student achievement based on aggregate state test scores when tested populations change. Each method uses longitudinal data to compare student performance in the current year (2021 in the example) to different historical norms. These adjustments are especially important for interpreting spring 2021 state assessment results. Regularly reporting these adjusted metrics could complement longitudinal descriptive statistics such as those suggested by An et al. (2022) to establish more valid interpretations of trends over time. This applies to nationally used interim tests as well, where directly comparing unadjusted results across years could provide misleading estimates due to changes in the populations of students taking tests.

All three methods lead to similar inferences about changes in student learning statewide during the pandemic when applied to Colorado assessment data. The adjusted differences were larger than differences based on directly comparing 2021 to 2019 scores, underscoring the

importance of using statistical adjustments to address confounding due to population changes. Relative to similar students in the prior cohort, tested students in 2021 made less progress in both subjects, with larger declines in math than ELA. The declines in scores were moderate and similar in magnitude to estimates based on applying these methods in other states.

These metrics have also been used to make causal inferences about the impact of the pandemic on student learning. While the adjustments support more accurate inferences about changes in student learning that can help understand pandemic impacts, they also face important limitations that test users should be aware of. First, statistical adjustments such as regression rely on strong assumptions about unconfoundedness to produce unbiased estimates of causal effects (Murnane & Willett, 2011). Second, these adjusted comparisons are more accurately interpreted as reflecting the effect of all experiences *during* the 2019 to 2021 period, and cannot be used to make inferences about the effects of specific pandemic-related disruptions or experiences (Bacher-Hicks & Goodman, 2021). Finally, whether they are used to support descriptive or causal claims, statistical adjustments can only be used to describe (at most) students who participate in testing and cannot completely solve challenges posed by missing data. For statistical adjustments that rely on prior test scores, the match rate statistics can be reported to: 1) remind audiences that not all students are represented by these analyses, 2) evaluate the generalizability of findings to the populations of interest, and 3) evaluate the sensitivity of results to missing data.

When statistical adjustments are used to compare progress across subgroups, decisions about the method of adjustment need to be aligned with intended uses. To estimate how factors specific to the past two years (which includes pandemic-related disruptions in 2021) impacted student learning, models that include demographic variables such as MR may be more

appropriate. From a policy perspective, however, if the goal is to promote equity by allocating resources where there appears to be the greatest need, in terms of where students made the least academic progress during the past two years (through a combination of pandemic-related factors and other pre-pandemic resource limitations), then the FT or SGP models may be more appropriate. In some contexts, there may also be limitations on whether adjustments based on demographics can be used. Policy and decision-makers should be clear about the policy goal of targeting resources and ensure that the statistical model used matches their policy aims (e.g., Ehlert et al., 2016).

Applying the types of statistical adjustments described in this paper to 2022 state data would require additional years of longitudinal data and be further reduced to higher grade levels, but many of the same considerations will apply to the interpretation of the results. States and other organizations should complement standard aggregate results such as average scores with metrics that anticipate and support intended interpretations and uses of the data. Finally, state or interim assessment data alone cannot provide information about how or why student learning was disrupted, or about impacts on other outcomes such as students' social and emotional well-being. Data from state assessments should be combined with information about other outcomes and students' OTL to more completely understand how students were impacted and to continue planning and monitoring recovery efforts (Marion, 2020).

References

- An, L. S., Ho, A. D., & Davis, L. L. (2022). Disrupted data: Using longitudinal assessment systems to monitor test score quality. *Educational Measurement: Issues and Practice*, 41(1), 28–32. <https://doi.org/10.1111/emip.12491>
- Bacher-Hicks, A., & Goodman, J. (2021). The Covid-19 pandemic is a lousy natural experiment for studying the effects of online learning: Focus, instead, on measuring the overall effects of the pandemic itself. *Education Next*, 21(4), 38–42.
- Balow, J., & Miller, C. M. (2020). *Restart & recovery: Assessments in spring 2021*. Council of Chief State School Officers.
<https://www.nciea.org/sites/default/files/publications/Assessments-Spring-2021.pdf>
- Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51. <https://doi.org/10.1111/j.1745-3992.2009.00161.x>
- Betebenner, D. W., & Van Iwaarden, A. (2022). *COVID-19 academic impact in Rhode Island*. National Center for the Improvement of Educational Assessment.
https://www.ride.ri.gov/Portals/0/Uploads/Documents/News/Rhode_Island_Academic_Impact_Report_FINAL.pdf?ver=2022-04-27-173245-260
- Betebenner, D. W., Van Iwaarden, A., Cooperman, A., Boyer, M., & Dadey, N. (2021, August 6). *Assessing the academic impact of COVID-19 in summer 2021*.
<https://www.nciea.org/blog/covid-19-disruptions/assessing-academic-impact-covid-19-summer-2021>
- Betebenner, D. W., & Wenning, R. J. (2021). *Understanding pandemic learning loss and learning recovery: The role of student growth & statewide testing*. National Center for

the Improvement of Educational Assessment.

https://www.nciea.org/sites/default/files/publications/CFA_LearningLossRecoveryGrowthStatewideTesting.pdf

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers. <http://www.ccsso.org/Documents/2013GrowthModels.pdf>

Cohodes, S., Goldhaber, D., Hill, P., Ho, A. D., Kogan, V., Polikoff, M. S., Sampson, C., & West, M. (2022). *Student achievement gaps and the pandemic: A new review of evidence from 2021-2022*. Center on Reinventing Public Education. https://crpe.org/wp-content/uploads/final_Academic-consensus-panel-2022.pdf

Colorado Department of Education. (nd). *Spring 2021 state assessment results: Interpretation considerations*. <https://www.cde.state.co.us/assessment/cmas-dataandresults>

Curriculum Associates. (2021). *Academic achievement at the end of the 2020– 2021 school year: Insights after more than a year of disrupted teaching and learning* [Research Brief]. Curriculum Associates. <https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/i-ready-understanding-student-needs-paper-spring-results-2021.pdf>

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy*, 30(3), 465–500. <https://doi.org/10.1177/0895904814557593>

Gewertz, C. (2021, October 21). State test results are in. Are they useless? *EducationWeek*. <https://www.edweek.org/teaching-learning/state-test-results-are-in-are-they-useless/2021/10>

Goldhaber, D., Kane, T. J., McEachin, A., Morton, E., Patterson, T., & Staiger, D. O. (2022).

The consequences of remote and hybrid instruction during the pandemic (Working Paper No. 30010). National Bureau of Economic Research.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177.
<https://doi.org/10.1111/j.1750-8606.2008.00061.x>

Ho, A. D. (2008). The problem with “Proficiency”: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, 37(6), 351–360.
<https://doi.org/10.3102/0013189X08323842>

Ho, A. D. (2021). *Three test-score metrics that all states should report*.
<https://scholar.harvard.edu/files/andrewho/files/threemetrics.pdf>

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960. <https://doi.org/10.2307/2289064>

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, 27(1), 3–17.
<https://doi.org/10.3102/10769986027001003>

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1–24.
<https://doi.org/10.1093/pan/mpr013>

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>

Keng, L., Marion, S., & Silver, D. (2022). *Using multiple achievement measures to understand the effects of COVID-19 on student learning*. National Center for the Improvement of

- Educational Assessment. <https://www.nciea.org/library/using-multiple-achievement-measures-to-understand-the-effects-of-covid-19-on-student-learning/>
- Kogan, V., & Lavertu, S. (2022). *How the COVID-19 pandemic affected student learning in Ohio: Analysis of spring 2021 Ohio state tests*. <https://glenn.osu.edu/how-covid-19-pandemic-affected-student-learning-ohio>.
- Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.
- Lewis, K., Kuhfeld, M., Ruzek, E., & McEachin, A. (2021). *Learning during COVID-19: Reading and math achievement in the 2020-21 school year*. NWEA. <https://www.nwea.org/content/uploads/2021/07/Learning-during-COVID-19-Reading-and-math-achievement-in-the-2020-2021-school-year.research-brief-1.pdf>
- Marion, S. (2020). *Using opportunity-to-learn data to support educational equity*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/sites/default/files/inline-files/CFA-Marion.OTL_.Indicators_0.pdf
- Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.
- National Urban League, League of United Latin American Citizens (LULAC), National Action Network (NAN), National Indian Education Association (NIEA), Southeast Asia Resource Action Center (CEARAC), UnidosUS, Council of Parent Attorneys and Advocates (COPAA), National Center for Learning Disabilities, National Center for Special Education in Charter Schools, The Education Trust, Education Reform Now, & Alliance for Excellent Education. (2020, November 20). *Open letter to Deputy Assistant*

- Secretary Ryder*. https://nul.org/sites/default/files/2020-11/Civil_Rts_Letter_to_DoED_re_Accountability_Guidance_w_signers_11192020.pdf
- Pendharkar, E. (2022, February 1). More than 1 million students didn't enroll during the pandemic. Will they come back? *EducationWeek*.
<https://www.edweek.org/leadership/more-than-1-million-students-didnt-enroll-during-the-pandemic-will-they-come-back/2021/06>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
<https://doi.org/10.1093/biomet/70.1.41>
- Shear, B. R. (2020). *Comparison of 2019 cohort and baseline student growth percentiles*. The Center for Assessment, Design, Research and Evaluation (CADRE).
<https://www.colorado.edu/cadre/node/373/attachment>
- Student, S. R. (2022). Vertical scales, deceleration, and empirical benchmarks for growth. *Educational Researcher*, 51(8), 536–543. <https://doi.org/10.3102/0013189X221105873>
- U.S. Department of Education. (2021). *Letter to chief state school officers*.
<https://oese.ed.gov/files/2021/02/DCL-on-assessments-and-acct-final.pdf>
- West, M. R., Lake, R., Betts, J., Cohodes, S., Gill, B., Ho, A. D., Loeb, S., McRae, B., Schwartz, H., Soland, J., & Walker, M. (2021). *How much have students missed academically because of the pandemic? A review of the evidence to date*. Center on Reinventing Public Education.

Table 1*Sample Sizes, Participation Rates, and Average Test Scores*

Variable	Grade 5 ELA	Grade 6 Math	Grade 7 ELA	Grade 8 Math
Sample Size				
2019	69247	69175	68725	67435
2021	64428	66816	68972	69086
Participation Rate				
2019	95.0%	93.5%	91.2%	87.3%
2021	72.8%	67.0%	62.3%	56.7%
Difference	-22.1	-26.5	-28.9	-30.5
Average Score				
2019	746.8	732.5	745.0	735.8
2021	746.1	726.6	742.0	730.3
Difference	-0.8	-5.9	-3.0	-5.4
Standard Deviation				
2019	34.1	31.0	38.5	41.2
2021	32.4	32.3	37.0	38.8

Note. The difference rows report the changes from 2019 to 2021.

Table 2*Comparison of Grade 5 ELA Enrolled, Tested, and Matched Populations*

Variable	Demographic Characteristics					Standardized Prior Scores		
	Enrolled	Tested		Matched		Enrolled	Matched	
	2021	2021	2019	2021	2019	2021	2021	2019
American Indian/Alaska Native	0.7%	0.6%	0.7%	0.6%	0.6%	-0.35	-0.24	-0.41
Asian	3.0%	3.0%	3.1%	3.0%	3.0%	0.34	0.37	0.34
Black	4.7%	3.8%	4.5%	3.7%	4.5%	-0.35	-0.29	-0.40
Hawaiian/Pacific Islander	0.3%	0.2%	0.3%	0.2%	0.2%	-0.39	-0.36	-0.12
Hispanic	34.8%	33.6%	35.1%	32.6%	34.4%	-0.33	-0.33	-0.37
White	51.6%	54.1%	51.7%	55.3%	52.7%	0.30	0.33	0.28
Two or More Races	4.9%	4.7%	4.7%	4.7%	4.6%	0.20	0.23	0.18
Female	48.9%	48.8%	48.6%	48.9%	48.9%	0.15	0.17	0.14
FRL	40.4%	38.6%	44.1%	37.9%	43.1%	-0.38	-0.37	-0.40
ELL	18.7%	18.3%	19.3%	16.5%	17.4%	-0.52	-0.52	-0.56
IEP	13.2%	11.9%	11.5%	11.8%	11.3%	-1.04	-1.02	-1.12
FRL (prior)				41.2%	45.0%	-0.37	-0.36	-0.39
ELL (prior)				16.8%	18.6%	-0.50	-0.51	-0.50
IEP (prior)				10.4%	9.6%	-0.92	-0.90	-1.01
N	64428	46917	65754	41532	57069	55541	41532	57069
Avg. Current Score	746.1	746.1	746.8	747.0	747.9			
Avg. Prior Score	740.0	741.3	738.8	741.3	738.8			
N Prior Score	55541	41532	57069	41532	57069			

Note. FRL=free/reduced price lunch eligible; ELL=state identified English Language Learner; IEP=student with individualized education plan. The FRL (prior), ELL (prior), and IEP (prior) variables indicate students' status relative to these programs two years prior (in 3rd grade) among matched students. The final three columns report average prior test scores, standardized relative to the full grade level distribution in 2017.

Table 3*Match Rates, Adjusted Differences, and Additional Reporting Metrics*

Variable	Grade 5 ELA	Grade 6 Math	Grade 7 ELA	Grade 8 Math
Prior Match Rate	68.3%	63.1%	59.4%	55.1%
Current Match Rate	64.5%	60.9%	56.6%	51.6%
Estimated Difference				
Observed (all)	-0.8	-5.9	-3.0	-5.4
Observed (matched)	-0.9	-6.0	-2.9	-5.8
FT	-2.5	-6.5	-5.3	-5.9
MR	-2.6	-6.7	-5.1	-6.0
MGP	-3.5	-11.8	-7.2	-9.6
Effect Size				
Observed (all)	-0.02	-0.19	-0.08	-0.13
Observed (matched)	-0.03	-0.19	-0.07	-0.14
FT	-0.07	-0.21	-0.14	-0.14
MR	-0.08	-0.21	-0.13	-0.15
Difference in % Proficient				
Observed (all)	-1.2	-5.4	-3.9	-7.4
FT	-3.6	-6.3	-6.5	-8.2
MR	-3.8	-6.6	-6.5	-8.3
MR Bounds				
LB	-0.05	-0.13	-0.08	-0.07
UB	-0.10	-0.30	-0.19	-0.22

Note. Observed (all)=observed (unadjusted) difference in average scores between all tested students in 2021 and 2019; Observed (matched)=observed (unadjusted) difference in average scores between 2019 and 2021 students in the matched samples. FT=fair trend difference; MR=multiple regression adjusted difference; MGP=mean baseline SGP; LB=lower bound of estimated effect described in text; UB=upper bound of estimated effect described in text. The FT, MR, and MGP estimates are based on the matched samples of students with current and prior test scores. Effect sizes are calculated relative to the standard deviation of scale scores in 2019 among all tested students.

Table 4*Adjusted Differences and Current Match Rates by Student Subgroup, Grade 5 ELA*

Group	N	Match	MR	FT	MGP	Ratio to Statewide Effect		
						MR	FT	MGP
All	41532	64.5%	-0.08	-0.07	-3.5			
Asian	1226	62.6%	-0.12	0.02	0.6	1.5	-0.2	-0.2
Black	1547	51.0%	-0.09	-0.17	-8.0	1.2	2.3	2.3
Hispanic	13523	60.4%	-0.11	-0.17	-8.3	1.4	2.3	2.4
White	22975	69.1%	-0.06	-0.01	-0.7	0.7	0.2	0.2
Male	21243	64.5%	-0.05	-0.10	-4.9	0.6	1.4	1.4
Female	20289	64.4%	-0.10	-0.04	-2.1	1.4	0.6	0.6
FRL	17120	65.3%	-0.09	-0.17	-7.8	1.2	2.2	2.2
Not FRL	24412	74.5%	-0.07	-0.01	-0.5	0.9	0.1	0.1

Note. Match=current match rate; MR=multiple regression; FT=fair trend; MGP=mean baseline SGP; FRL=free/reduced price lunch eligible in grade 3 in 2019. MR and FT estimates are in effect size units relative to the standard deviation of scale scores in 2019 among all tested students. The ratio columns are computed using unrounded estimates and differ slightly from the ratios implied by the rounded estimates reported here.

Appendix

A single-step multiple regression (MR) model to estimate the effect of the pandemic while adjusting for prior scores and student demographics could be specified by combining longitudinal test scores from both the 2019 and 2021 cohorts. A model of the following form would be estimated for each grade and subject:

$$y_{igy} = \beta_0 + \beta_1 \text{spring2021}_{iy} + \delta \mathbf{Z}_{i(g-2)(y-2)} + e_{igy}. \quad \text{Eq. A1}$$

Here y_{igy} is the test score for student i in grade g in year y , spring2021_{iy} is an indicator variable equal to 1 for observations in spring 2021 and 0 for observations in spring 2019, and $\mathbf{Z}_{i(g-2)(y-2)}$ is a vector of covariates from grade $g - 2$ and year $y - 2$ that includes demographic variables and the lag-2 prior test scores. In this model the coefficient β_1 represents the average difference in scores between students in the 2021 and 2019 cohorts, adjusting for differences in the variables included in \mathbf{Z} , and would serve as the estimated effect of the pandemic when making a causal inference about the pandemic effect. The specific variables included in \mathbf{Z} can vary and it would be possible to include data from additional pre-pandemic cohorts.

This model cannot be used to directly produce estimates of differential effects across demographic subgroups. There are two ways to extend the model to estimate differential effects across subgroups: 1) include interaction terms, or 2) estimate the model separately for each demographic subgroup of interest.

Adding interaction terms to the model in Eq. A1 can be used to produce identical estimates to the two-step MR estimates described in the main text. This proceeds by estimating the following model that interacts the *spring2021* indicator with all other covariates (including prior scores) and centering all covariates relative to the 2021 cohort means, $\bar{\mathbf{Z}}^{(g-2)(2019)}$:

$$y_{igy} = \beta_0 + \beta_1 \text{spring2021}_{iy} + \delta(\mathbf{Z}_{i(g-2)(y-2)} - \bar{\mathbf{Z}}^{(g-2)(2019)}) \\ + \gamma \text{spring2021}_{iy} (\mathbf{Z}_{i(g-2)(y-2)} - \bar{\mathbf{Z}}^{(g-2)(2019)}) + e_{igy}. \quad \text{Eq. A2}$$

In this model, β_1 estimates the average difference between observed scores in the 2021 cohort and the scores we would expect for these students based on students with the same prior scores and demographic characteristics in the 2019 cohort. This will be identical to the statewide average difference MR_g in Eq. 6. Average marginal effects estimated among students in the 2021 cohort for each of the demographic subgroups represented by covariates in \mathbf{Z} will also be identical to the MR_g differences estimated within each subgroup and reported in Table 4 (Imbens & Rubin, 2015).

Estimating effects separately by subgroup produces conceptually similar estimates, but they will not necessarily be identical to those reported in the main text. In the CO data, for example, subgroup effects estimated by fitting the model in Eq. A1 separately for each subgroup listed in Table 4 produced estimates that were within 0.003 of the values in Table 4. Differences between the two approaches at other grade levels were similar.

Table A1 reports the regression model parameter estimates for the MR model based on data from the 2019 cohort used to calculate the MR adjusted differences in the main text.

Table A1

Regression Model Estimates Predicting 2019 Scores From 2017 Prior Scores and Demographic Variables

<i>Predictors</i>	Grade 5		Grade 6		Grade 7		Grade 8	
	<i>Est</i>	<i>SE</i>	<i>Est</i>	<i>SE</i>	<i>Est</i>	<i>SE</i>	<i>Est</i>	<i>SE</i>
(Intercept)	0.060	0.005	0.024	0.005	-0.014	0.005	-0.028	0.005
Prior Score	0.798	0.005	0.861	0.004	0.785	0.005	0.891	0.004
Prior Score^2	0.011	0.002	0.071	0.002	0.041	0.002	0.072	0.002
Prior Score^3	-0.027	0.001	-0.031	0.001	-0.027	0.001	-0.038	0.001
Female	0.114	0.005	0.026	0.004	0.174	0.005	0.060	0.005
American Indian or Alaska Native	-0.085	0.032	-0.027	0.028	-0.097	0.032	-0.073	0.029
Asian	0.073	0.015	0.116	0.014	0.158	0.016	0.139	0.013
Black	-0.104	0.013	-0.110	0.011	-0.133	0.013	-0.076	0.012
Hawaiian/Pacific Islander	0.036	0.053	-0.158	0.045	-0.116	0.057	-0.047	0.048
Hispanic	-0.100	0.007	-0.098	0.006	-0.102	0.007	-0.077	0.007
Two or More Races	-0.022	0.012	-0.015	0.011	-0.005	0.013	-0.007	0.012
FRL (prior)	-0.133	0.006	-0.125	0.005	-0.148	0.006	-0.087	0.006
ELL (prior)	0.046	0.008	0.032	0.007	0.062	0.008	0.051	0.007
IEP (prior)	-0.202	0.009	-0.170	0.008	-0.182	0.010	-0.178	0.009
N	57069		57377		55052		51605	
R ²	0.65		0.72		0.63		0.72	

Note. Current and prior scores standardized within grade. The FRL, ELL, and IEP variables are designations at the time of the prior score, 2017.